How Computer Science Helps Feed the World

Bruce Yellin
EMC
bruce.yellin@emc.com

Stephen Yellin
Independent Communications Consultant
StephenYellin@yahoo.com
# Table of Contents

Introduction ........................................................................................................................................... 3  
What is the Food Supply Chain? ........................................................................................................... 3  
   The Food Supply Chain ....................................................................................................................... 4  
Computer Science on the Farm ............................................................................................................. 6  
   Geotagging, Telematics, Precision Farming and Controlled Traffic Farming ..................................... 6  
   Robotics and Automated Milking ........................................................................................................ 8  
Big Data on the Farm ........................................................................................................................... 11  
   Farmeron and Agrivi - Platform 3 Apps that Deliver Real-time Big Data ........................................ 12  
   Badger-Bluff Fanny Freddie – Big Data and Best of Breed ............................................................... 14  
   Seeds, Crop Insurance and Big Data .................................................................................................. 16  
   Big Data That Flies ........................................................................................................................... 17  
Computer Science and the Manufacturer ............................................................................................. 18  
   ConAgra and Peanut Butter – A Visit to the Manufacturer ............................................................... 19  
Computer Science and the Distributor ................................................................................................ 24  
   The Food Supply Chain Depends on Trucking ................................................................................. 28  
Computer Science in the Grocery Store ............................................................................................... 33  
   The Barcode “Miracle” ....................................................................................................................... 33  
   How does the Point of Sale Terminal work? ..................................................................................... 35  
Data Warehousing and Big Data .......................................................................................................... 37  
   Data Warehouse Basics .................................................................................................................... 38  
   Hadoop Basics .................................................................................................................................. 42  
Conclusion ........................................................................................................................................... 45  
Appendix – List of Abbreviations ........................................................................................................ 47  
Footnotes ............................................................................................................................................... 48  

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect EMC Corporation’s views, processes or methodologies.
Introduction

Food is essential to life. Our ancestors originally hunted and gathered food, plucking fruit from a tree, throwing a spear at an animal, or even clubbing a fish. Eventually humans learned to grow crops and breed livestock, not just hunt, marking the birth of agriculture and civilization. As time went on, the number of people needed to sustain everyone else began to decrease. The expansion of science and technology beginning in the Industrial Era greatly reduced the number of people actively involved in creating food. Technological innovation in machinery and methods has allowed society to advance where only a minority is needed to feed the rest of the planet.

These technological breakthroughs, together with other improvements such as the motorized combine have made it possible to feed a global population of billions\(^1\). As the global population has more than doubled from 1961 to 2012, the United Nations Food and Agriculture Organization shows that various food groups have maintained or exceeded the population’s growth rate as farmers attempt to feed the world. While food distribution still limits the ability to feed everyone properly, crops like sugar cane have increased 4-fold, wheat has tripled, soybeans have increased 9-fold, and the metric ton production of cow milk has doubled.

Yet new innovations will be needed to meet an ever-increasing demand for food. Today over 7.5 billion people inhabit Earth, and by 2050 that number could approach 9-10 billion\(^2\). Along the way, the number of farms is expected to continue to decline in favor of large, corporate-run entities that perform the same function. Computer science will be critical in pointing the way forward. This is what makes studying the role of computer science in today’s agricultural process so important and why we have chosen to write about it.

What is the Food Supply Chain?

Henry Ford’s 1913 Model T assembly line is an early example of how a supply chain helped make cars affordable\(^3\). To feed the world’s population in an efficient and profitable manner, food production must follow a multi-sourced food supply line process.
that integrates all facets of this illustration. Most supply chains follow similar paths, such as producer \( \rightarrow \) manufacturer \( \rightarrow \) distributor \( \rightarrow \) wholesaler \( \rightarrow \) retailer \( \rightarrow \) consumer. Rewards like money flow in the opposite direction. To keep the chain optimized, information must also flow.

With food, there are many interactions just at the farm phase alone. For example, there are the sales of heavy equipment, telephony and communications gear, computer equipment, apparel, insurance, electricity, gas, government permits and regulations, real estate transactions, and many more. Depending on the farm product, veterinary service, research and consulting, and feed production are involved. At the production phase, the meat packing can be labor-intensive, and many forms of packaging are required along with the plant infrastructure itself, which again needs equipment, power, technology, etc. Connecting the farmer to the plant requires trucks or other transportation, workers, roads and infrastructure, police, etc. The amount and types of detailed interaction that goes on with a supply chain such as food is almost inconceivable. Once the product is produced the next phase of food distribution kicks in with more of the same interactive complexity as the process moves one step closer to the grocery store and the consumer. The grocery store accepts deliveries from the distributor in a just-in-time (JIT) manner as there is little back room storage at the store – most of the space is dedicated to shelves and customers. As shelf inventory runs low or in advance of anticipated demand, additional quantities are ordered, which triggers the distributors to put more goods on trucks. Communications along the food supply chain (FSC) keeps all players aware of consumer demand so production can ramp up or down as the entire market dynamics runs 24 hour a day. Throughout the chain, tracking food throughout the processes is critical to food safety. At the end of the cycle is waste disposal and recycling.

**The Food Supply Chain**

The process of taking food from “farm to fork” is a complex chain of steps that stretches from the creation of the product to its consumption by consumers all over the globe. The FSC process varies from foodstuff to foodstuff, but each product’s supply chain shares common characteristics with the others. In order to illustrate the essential role of computer science in today’s FSC, let’s examine the steps of the chain themselves.

- **Growing/Rearing** - A farmer decides to grow certain crops or livestock, usually based on market demand. Crops such as grains, fruits, and vegetables are grown and cultivated from planted seeds, while livestock is bred then reared on the farm.
- **Harvesting** - Crops are collected from the fields, either by harvesting grains and vegetables from the ground or picking fruits from a tree. Livestock is either chosen for slaughter or, in the case of cows, milked either by hand or by automated machines.
• **Storing** - Foodstuffs are stored in different containers or silos, varying in size and temperature based on the product being created. For example, milk is stored at a minimum of 39°F for 48 hours, and with ice cream, the milk is frozen and not chilled.  
• **Processing/Testing** - Foodstuffs are taken to a processing plant to convert the raw material into products for public consumption. While at the factory the raw material is tested to ensure it meets legal quality standards.  
• **Packaging/Selling** - Workers package most products, adding required consumer information such as expiration dates, ingredients and calorie counts. Finished products are sent to a warehouse for distribution to grocery stores, restaurants, etc.  
• **Consumption** - Customers purchase and consume the finished products.

Throughout this process, trucks or other transport are employed to deliver goods to the right location, on time, in sequence, and at the right price. Items may require packaging, internal movement on conveyor belts, forklifts or overhead monorails, storage, and retrieval. Items then may need to be sorted, picked, prepared for shipment, and dispatched to outgoing transports.

All this activity is closely monitored through computer science methods employing concepts like barcodes, radio frequency identification tag (RFID), magnetic strips, camera vision, and more.

The farmer, manufacturer, warehouse, and store are highly dependent on transportation for each step, data to relay quantities, locations, and timing, all while minimizing overall costs and adhering to regulations to deliver the food to the consumer. This seamless, complex interaction of renewable and non-renewable resources (such as the people, ground, water, air, seed, fossil fuels, solar and wind power) involves whole foods as well as the creation and formulation of ingredients that are key to what the world consumes. No single stage in this process operates in isolation. For example, the farmer needs equipment, workers, fuel, seed, water, sunlight, and more, and consumers interact with the grocer in terms of workers, selection, payment, parking, etc. Some food follows a direct path as when the consumer visits a farm to pick strawberries or sticks a fishing line in the ocean.

Consumers also get food from service industries, like one of McDonald’s 35,000 restaurants that interact with utilities, governments and the media, in addition to their own labor force and a supplier chain that buys food and ingredients. Computer science is fundamental to this entire set of supply chains.

The FSC needs continuous feedback to allow the farmer to gauge market conditions, the manufacturer and distributor to forecast market demand, and the retailer to know what
customers like to buy. With approximately 40% of the world’s population involved in agriculture, data plays a critical role in the FSC\(^5\). The many variables in the FSC must all come together to optimize production and reduce costs, all the while adjusting to a decrease in the number of farmers and farms – a big data problem. The FSC is focused on the grocery store where maintaining JIT delivery and high volume are critical to this low profit margin industry\(^6\). Computer science helps this chain reduce risks and costs, increase yield, streamline complex logistics, and help provide more nutrition to the world. Let’s start with the farm.

**Computer Science on the Farm**

Farming is historically labor intensive and involves a fair amount of risk. Farmers need to monitor and react to the weather as well as the market for their products. They must optimize yield, reduce waste, maintain food safety, and understand the environmental impact, supplier interaction, and product delivery along the FSC. Let’s explore three areas where computer science helps the farmer improve productivity and greatly lowers risk – telematics and geotagging, automated milking, and big data seeding.

**Geotagging, Telematics, Precision Farming, and Controlled Traffic Farming**

Geotagging is a process of adding latitude, longitude and other geographical metadata to an object such as a tractor. Telematics is a combination of telecommunications and informatics that gives a machine geotagged location information, fuel consumption, vehicle safety, quantity of crop harvested, and more, helping a farmer optimize yield and lower expenses. Precision farming takes into account that crops have non-uniform needs based on variability of the fields, and by using Global Positioning System (GPS) navigation, geotagging and telematics, it can "...apply the right treatment in the right place at the right time..."\(^7\). With transponder-equipped farm machinery, an interactive map can be established with space-based accuracy that delineates areas of a field that need more (or less) water, pesticide, fertilizer, etc. Together, these technologies increase crop production, improve profitability, and conserve water and chemicals because they are focused just on the crops that need them. Equipment such as a John Deere combine equipped with JDLink Telematics can have its mechanical health, fuel utilization, maintenance status, operator alerts, and the number of machine hours logged\(^8,9\).

The computer science that enables the GPS and Russian Global Navigation Satellite System (GLONASS) is really a modern marvel. A network of satellites equipped with synchronized atomic clocks constantly broadcasts their current time and position. GPS employs 32 satellites
while GLONASS uses 24\textsuperscript{10}. A GPS receiver picks up the transmission from a minimum of four satellites 12,600 mi above Earth to compute the distance to each satellite\textsuperscript{11}. Radio signals travel at the speed of light (186,000 MP/s), so a farmer can determine where GPS-enabled equipment is from a desktop application that displays the time it takes to receive the transmission divided by the speed of light. The distance from the four satellites and a high school math concept called “trilateration” gives you the data points to let you know exactly where you are.

To illustrate how it works, here is a simple two--dimensional trilateration problem. Suppose you were lost and someone says you are 625 miles from Boise, Idaho. Good information, but all you would know was that you were somewhere along this circle 625 away from Boise\textsuperscript{12}. If you were then told you were also 690 miles from Minneapolis and 615 miles from Tucson, trilateration would tell you are in Denver. This concept also works in three dimensions enabling a farmer to find their GPS-enabled tractor based on satellite transmissions.

This science all culminates in the discipline of Controlled Traffic Farming (CTF). Equipment can have sets of massively wide tires that crush plants, and in excess of 25% of a field can easily be trampled\textsuperscript{13}. Land is needlessly compressed and unable to grow crops when a farmer manually steers a tractor and tries to drive in an absolutely straight line. Technology has been developed to minimize the compression by allowing the machinery to be driven by a computer in the driver’s cab. The computer takes in telemetry signals and precisely guides the machine to follow the exact same path it did the last time it circumnavigated this field. With equipment path tracks precisely and permanently chosen, and all heavy machinery using the same track width, field yields have increased by an average of 12% from the same amount of land\textsuperscript{14}. As yields increase, less seed is needed, fertilizer and pesticide use is reduced, water seeps into the ground easier, and erosion is reduced. Fewer hours are also placed on the machinery, allowing the farmer to spend less time in the fields and save as much as 30-50% in fuel costs. As one farmer noted, before CTF, they needed 120 liters of diesel fuel per hectare (about 2.5 acres). Using CTF cut their fuel need in half\textsuperscript{15}.
What makes this approach feasible is “autosteer” machinery enabled by GPS and Real Time Kinematic (RTK) systems. While standard GPS signals are suitable for tillage, planting, and pre-emergence spraying, RTK-enabled equipment can operate within “…one inch pass-to-pass and year-to-year” of a perfect route\textsuperscript{16}. RTK enhances the GPS accuracy by augmenting the satellite signal with an earth-bound fixed point base station that transmits tiny differential corrections.

Crops harvested by a combine need a grain truck to capture the product. With GPS/RTK operated equipment, the repository truck can be autonomously driven next to the combine, mirroring the combine’s path without the need for an operator. As a truck is filled, another can automatically be brought into play. For example, Kinze Manufacturing created an Autonomous Harvest System that uses a driverless tractor and a computerized grain cart to work with a human combine driver\textsuperscript{17}.

Autonomous Tractor Corporation created a robot tractor that can operate 24x7 precisely following a path and taking the burden out of farming while dramatically increasing productivity. A farmer “trains” the tractor what to do and it repeats those steps based on a laser-guided Area Positioning System (APS) rather than a GPS system.

**Robotics and Automated Milking**

For most of human history the process of getting milk from the dairy farm to the kitchen table has begun with long, hard hours of a farmer and his helpers milking cows by hand. Keeping reliable employees that will show up whenever the cows need milking – each cow’s “milking speed” differs based on their physical activity and consumption on a given day – is costly, and the work of milking a cow by hand, tedious\textsuperscript{18}. Sickness and insufficient production means cow herds often have to be culled, making it hard for dairymen to turn a profit in today’s milk market.

Milk is vital to feeding the world over the next few decades. Production of milk is expected to grow 19% by 2020, from 692 million tons to 827 million tons, while the number of dairy farms has declined 30% over the last decade and over 60% in the last thirty years\textsuperscript{19,20,21}. In the United
States, almost 60% of milk production comes from large dairies of 500+ cows. In general, a cow produces 8 gallons/day (about 70 pounds of milk)\(^{22}\). The most productive cows are found in Saudi Arabia, Israel, and the Republic of Korea. A cow in the U.S. produces over 21,000 pounds of milk a year.

Since 1944 the U.S. cow population has shrunk by two-thirds from 25.6 million cows to 9.3 million cows, but produces 59% more milk thanks to technological improvements\(^{23}\). Automated Milking Systems (AMS) were introduced in the late 20\(^{th}\) century as automation and computer science began replacing human labor in the milking process\(^ {24}\). According to the U.S. Department of Agriculture's National Agricultural Statistics Service, “… the average annual milk production per cow in 1950 was 5,314 pounds. By 2000, it had more than tripled to 18,204. By 2011, as the graphic shows, it had topped 21,000”\(^ {25}\).

**How does it work?**

When “Elsie” the cow feels the need to excrete milk, she follows the training she was given and walks into a large barn and up to an automated gate. If the milking stall is empty, the AMS detects Elsie’s unique transponder which she is wearing around her neck, and since the database shows she has not been milked in a few hours it lets her in. Cows coming in after Elsie naturally queue up until she is done milking. If the machine determines Elsie has been milked too often, she is guided through the stall without being milked and not given a treat.

Once in the stall, Elsie starts eating a specific amount of her favorite corn meal or roasted soybeans pellets\(^ {26}\). The AMS knows where Elsie’s teats are located based on past milking history. As an added precaution, the robot may use ultrasound, laser sensors, or 3-D camera imagery to detect, accurately scan, and map her udder for reference points for the robotic arm. The robot then uses this Elsie-specific
data to move a computer controlled mechanical arm equipped with teat cleaning brushes from outside the stall to underneath Elsie, and cleans her four teats. After cleaning, the robot then attaches the teat milking cups to Elsie one at a time based on the mapped coordinates.

This basic flowchart from patent EP0360354A1 illustrates the steps and decisions the AMS must take in instructing the laser beams to gather information, guide motor actuators to move to the correct coordinates to ensure the equipment hooks up to Elsie properly, and record data about the event. The AMS then starts the milking and turns off when the milk flow stops. In essence, Elsie is milking herself. In the photo of Lely’s Astronaut A4 AMS on the previous page, you can see how the robot arm has positioned itself under the cow’s udders. When the milking ends, the robot retracts the teat attachments, cleans Elsie’s teats, and opens a gate for her to leave.

The robot steam cleans the attachments in preparation for the next queued cow. Her milk is tested, and data is sent to a central computer and database with information on flow rates, elapsed time, milking speed, milk quality, how much each cow has eaten, their weight, milk fat content, and even how many steps it took today. AMS is said to improve a cow’s health with a reduction in instances of mastitis (inflammation of the cow’s udder). At any time, the farmer can get a detailed report on Elsie using their smartphone.

Cows are individually milked by AMS as many as 5-6 times a day, 24 hours a day with each machine milking 60-105 cows a day\textsuperscript{27}. One robot can also serve two cows at the same time when the system is placed between two stalls. The latest systems are circular allowing for more robots and thus more cows are milked simultaneously\textsuperscript{28}. 

\textsuperscript{27} Whether the number of milking times per day is 5 or 6 is not clear, but it could be both. 
\textsuperscript{28} Whether the number of cows milked per day is 60 or 105 is not clear, but it could be both.
AMS brings significant benefits to the dairy farming industry. For people like Tom Borden, a sixth-generation New York dairy farmer, AMS is the difference between higher production with much less physical labor, or going out of business. “Either we were going to get out, we were going to get bigger, or we were going to try something different”, said Borden, who like many dairy farmers is now trying “something different”.

The AMS increases efficiency, reduces costs (especially labor costs which often includes room and board, insurance, etc.), improves product quality, reduces environmental impact, and improves animal health and welfare. With an AMS, the farmer no longer needs to get up at 4AM to begin milking, nor do they have to continue working until 10PM at night. As farm labor costs rise, AMS naturally looks attractive. This chart shows the labor gains through automation as milk production increases faster than its labor component.

AMS also has some drawbacks, however. At $200,000-$250,000 a machine plus the potential cost of redesigning the barn where milking is done, AMS is expensive. The Bordens, for example, spent $1.2 million on 2 machines and rebuilding their barns to accommodate them. A greater reliance on machinery also means increased operational costs. These startup costs may prevent small-business dairy farmers from purchasing AMS systems.

Each cow’s milk is held in the farm’s giant chilled storage tank. After the tank is emptied into a milk truck, it is brought to the manufacturer/producer where the milk undergoes further testing; the milk fat is separated, homogenized, pasteurized, and chilled again. It is then packaged, labeled, and dated with an expiration date, then stored or loaded on a distribution truck to the distributor or grocery store. If the manufacturer needs butterfat for other products, it is separated from milk using a centrifuge. If they are using the milk to make yogurt, they pasteurize and homogenize the milk, then chill it, add culture and perhaps some fruit or flavoring, and package it.

**Big Data on the Farm**

There is a mountain of raw data coming out of the FSC with much of it originating on the farm. The data is from the fields, barn, and the tractor, and is critical in helping to feed the world over the next few decades. Farmers want to analyze the data to improve productivity and profitability. Seed, fertilizer, chemical, and equipment companies use data for product development and
placement, and increased profits. Land owners, insurance companies, and financial institutions access data for risk assessments and pricing. From automated soil analysis that leads to soil composition decision making, real-time actionable data by the farmer, breeding cows that give the highest milk production, financially protecting the farmer from droughts, to making better decisions with seed and crop sensing, big data promises some incredible farm innovation.

**Farmeron and Agrivi - Platform 3 Apps that Deliver Real-time Big Data**

Imaginary farmer “Steve” leads a hectic life. He enjoys getting his hands dirty in the field, but he also knows he has to get them even “dirtier” in his farm’s data. In addition to growing vegetables Steve has administrative duties related to hiring additional workers at harvest time, tracking the weather, creating farm progress reports, working with an agronomist (a specialist in soil management and crop production), and meeting with his business coordinator and accountant. Steve has basic sales skills and knows he needs to be more computer-savvy. Steve’s answer is a “platform 3” (cloud) application\(^\text{34}\). They represent an easy, graphical way to gain valuable insights into his online farm data and other information so he can produce actionable real-time metrics and recommendations, all through his smartphone, tablet, or personal computer.

**Farmeron**

Matija Kopić is the CEO of the four year old Croatian-based Software-as-a-Service (SaaS) cloud computing company called Farmeron, which brings a powerful yet easy to use platform 3 app to a farmer’s smart device. It allows Steve the farmer to keep all his dairy data in one place free of the complexity of running his own farm management system on the premises. The system aggregates aspects like feeding and milking to allow for better decision making, leveraging a knowledge base that encompasses over 1,000 farmers. It tracks partners, physical plant infrastructure like the farmer’s barns and equipment, feed orders, and more.

One farmer found that just 5 months of Farmeron use helped improve milk production by 34%. Another farmer was able to record herd events in less than 5 minutes in contrast to an older, time-consuming set of complex Excel spreadsheets\(^\text{35}\). Farmeron analytics features include “…an intelligent weighing system that lets farmers track animal weights against feed mix in order to optimize diets. Farmers also can log medical records for each animal, and get insights into cow breeding cycles and milking, among many other uses.”\(^\text{36}\)

Farmeron can display a herd’s conception rate, number of calvings, milk protein, fat, Somatic Cell Count (SCC shows milk quality), lactation yield, and other trends, or drill-down into a cow's
health and medical history, diet, and milk production, and tie all that into a farmer’s finances. The app interfaces with RFID readers and leverages a smart device’s Bluetooth to record data. It also has cloud-based analytics and simulations of likely outcomes of management decisions. "For example, you can see what would happen if you add 50 to 100 cows to your operation or if you made a significant change in your feeding program." It can also answer big-data questions like “How will an increase in corn prices impact my farm operations?” and “If my cows eat an additional pound of food, how much more milk production can I expect?”

Behind the scenes, Farmeron administrators are doing the “heavy lifting” of database design, defining procedures, scripting, database administration, optimization, configuring software, replication, and more. As shown on the right, their cloud interfaces with many external systems such as the Lely AMS and other farm activities.

Farmeron’s web-based dashboard provides real-time data about a farm, even down to the level of each individual cow. The data is presented in bright graphs and charts. This cloud-derived chart shows herd and fertility metrics and information on a herd’s productivity in contrast to their gynecological status.

The service costs 25 cents a cow per month for farms with up to 75 cows, or 45 per cow/month in a 600 animal herd. They estimate the farm management software market at over $12B annually as 14 million medium-sized and larger corporate farms spend $900 annually on software costs, together with 150 million small farms who may or may not use software.

**Agrivi**

Another Croatian startup, Agrivi, has a platform 3 app that focuses on cloud-based farm management with an integrated knowledge base “…for over 60 crops.” The app coordinates tasks and expenses, such as the purchase and application of fertilizers and pesticides, expenditures for fuel, and logging work hours. When products are bought or sold, they are entered into the app. Agrivi keeps the inventory, provides weather forecasting information and history for the fields, and alerts concerning plant disease or pests. All of this is
through a visual farm manager dashboard on the farmer’s smartphone or tablet, helping alleviate the farmer’s bottlenecks and increasing profitability.

Using a farmer-task approach, Agrivi uses a project approach to simplify and make quick work of planning, monitoring, and tracking farm activities and other variables such as fuel consumption, fertilizer and pesticide application, inventory with alarms, and hours worked. The system easily generates reports focused on:

- Farm management based on expert templates
- Field utilization and efficiency
- Inventory tracked to prevent waste and aid in reordering
- Finances covering what was sold and what was purchased
- Full and part-time workers, and their assigned tasks
- Maintenance and fuel consumption
- Calendar for personal and company tasks
- Details on each field

The system keeps the farmer aware of the weather with 7-day forecasts, and comparisons to the past 3 years are made readily available to leverage disease risks. In the field, the farmer can tap into a knowledge base of best-practices and access all the necessary documents since they are stored in the cloud. Data is kept in secure data centers with daily backups. The app is downloadable from the iTunes app store or Google Play, and it is free for one farm and one user (see chart at right). The pricing for additional users and farms is very straightforward. No contract is required and paid plans can be charged to a credit card or PayPal.

**Badger-Bluff Fanny Freddie – Big Data and Best of Breed**

Cows produce milk when they are lactating, so they must be bred and give birth to calves. Farmers often use artificial insemination to keep the cycle productive. They have found that they can selectively breed dairy cows that excel at milk production. Recent studies have shown that selective breeding can boost milk production by 8%, and that 500 U.S. bulls are used to inseminate 9 million cows.

Over the last few years, the amount of livestock data has skyrocketed – from genetics, real-time RFID inputs with data from automated milking systems and meters that record milk yield and
speed, temperature, fat and protein concentrations, and other data like eating habits and exercise. Most of the data originates on the farm and contains animal identification, production environment, and performance, with a lot of it stored in a national dairy database\textsuperscript{49}.

The biggest issue is how to analyze it and create actionable results. Data mining has turned out to be one of the primary use cases for this big data challenge. The goal is to perform cluster analysis, data classification, and regression analysis to discover data relationships. Hadoop plays a big part in processing the huge amount of raw data. As Cole \textit{et al.} write, classification models like these are easily transformed by computer programs like MapReduce to accurately predict data such as breeding values, feed intake or milk yield\textsuperscript{50}.

Key to increasing milk production is the use of computer science and big data to select the appropriate genes that help females produce milk, out of a cow’s genome of 3 billion base pairs and 22,000 genes. In 2009, the USDA examined the 50,000 markers of a Holstein bull’s DNA that led to the creation of a bull named Badger-Bluff Fanny Freddie. This bull’s semen is in big demand for artificial-insemination. Before big data, a bull named Pawnee Farm Arlinda Chief had been recognized as the best for milk and protein production with over 16,000 daughters, 500,000 granddaughters, and 2 million great granddaughters\textsuperscript{51}. So naturally the expectations for Badger-Bluff Fanny Freddie were dramatically higher.

Similar to baseball statistics that assign a value to a player’s total contributions to a team (Wins Above Replacement), the key statistic used to compare a bull’s real net worth is a single dollar value called lifetime net merit (LNM). A bull that helps a cow produce an extra 1,000 pounds of milk in her lifetime has a LNM of $1\textsuperscript{52}. Badger-Bluff Fanny Freddie sets the world’s record with a LNM of $792\textsuperscript{53}. John Cole, an animal scientist, estimates through genetic fine tuning and big data analysis that the next “Freddie” could be 10 times “better” than Freddie with a LNM in excess of $7,000. Techniques like this will boost milk production and help feed the world, even with a smaller dairy herd.
Seeds, Crop Insurance and Big Data

Big data-based farming is a big business. Monsanto, the seed and biotech giant with annual sales of $16 billion, set a new direction for its company as it sought to leverage big data by providing its customers with “…agronomic practices, seed genetics and innovative on-farm technology to deliver optimal yield to farmers while using fewer resources.” Monsanto sees a market for delivering customized reports and products tailored to a farmer’s soil conditions. As a result, they acquired Precision Planting in 2012 and The Climate Corporation in 2013.

Precision Planting was known for improved seed spacing, better depth control, and producing plants with better root systems. The Climate Corporation, an agriculture data sciences company, was founded by Google data scientists. Brought together, they introduced Monsanto FieldScripts to tackle the problem that “Corn yield varies across a field. Even in a good year, there are above- and below-average sections of a field. Varying the seeding rate across the different sections of the field may increase yield potential for farmers.” FieldScripts allows for the planting of “…less seed in low potential areas and more seed in high potential areas.” A two-year test shows corn yields increase by 5% or 5 to 10 bushels per acre. A 10 bushel increase per acre, at $6/bushel on a 2,000 acre farm, could boost the farmer’s income by $120,000. Field “prescriptions” are stored in the FieldScript cloud, and when the farmer connects their iPad to a Precision Planting seed drill pulled by a GPS guided tractor, varieties of specially engineered Monsanto seeds are precisely deposited in the correct proportions, spacing, and depth according to The Climate Corporation’s big data weather analysis. FieldScripts is expected to help increase field yield by 25%. They charge $10 per FieldScript acre and believe the market to be in excess of 1 billion acres.

This precision agriculture – using computer science technology – goes a long way to help feed the growing population.

The Climate Corporation, with over 200 scientists on staff, also insures farmers against the loss of profits because of bad weather. Their mathematical data scientists, agronomists, and climatologists build advanced data models that deliver valuable field-level information to the farmer, leveraging the public cloud computing model. They use Amazon Web Services’ Elastic
Compute Cloud (EC2) for server compute power, Elastic MapReduce (EMR) to process the data, and store information on the Amazon Simple Storage Service (S3). Using EMR, analytics, and modeling, they issue insurance policies for freezes, daytime and nighttime heat stress, drought, and heavy rain. The same models help The Climate Corporation simulate weather patterns and understand their risks should bad weather result in farmer-initiated claims. “For each point on a grid, it has come up with 10,000 scenarios that could affect a grower two years out.” Eventually they expect to be able to model 100,000 such scenarios. Their model uses 60 years of crop data, 14 TBs of soil data, and 1 million government Doppler radar points, all processed by 50 EMR clusters every day. They also batch process 20-30 TB data sets a month that can run for days and use thousands of instances of AWS’s tuned version of MapReduce. The business’s asymmetric nature is well suited to the public cloud.

**Big Data That Flies**

Farmers generate an incredible amount of data every year through the care, feeding, protection and harvesting of their crops. When dynamic conditions are factored in, such as the weather, crop yield, market fluctuations, machinery upkeep, soil type, soil moisture, soil nutrients, pest conditions, and satellite and drone data, it is difficult to make sound business decisions. Simply knowing what seed to plant in a specific field becomes complex as different varieties exhibit specific genetic traits, such as their water consumption and resistance to insects. The vast amount of information doesn’t stop there – companies that specialize in custom seed, fertilizer, chemicals, and machinery use data to develop and market new products, while land owners, financial institutions, and insurance companies base risk assessments and pricing on big data.

In the past, farmers optimized farm activity using satellites or manned aircraft to collect images. A plane could cost $1,000 an hour to photograph a field. For under $1,000, they can use an airplane or helicopter drone to accomplish the same mission. Drone supplier examples include:

- **3D Robotics X8-M** - can map 25 acres and fly for up to 14 minutes. With a 12 MP camera and image processing software, it creates georeferenced and orthorectified maps based on a farmer’s specified route.

- **PrecisionHawk** - automatically surveys a field with 12 different sensors including a camera, multispectral, thermal, LiDAR, and hyperspectral. It can stay aloft for an hour and produce a 3D terrain map showing plant height, weeds, and plant counts, and compute a “health
The drone uses an embedded 600 MHz processor running Dronecode open-source Linux and cloud-software to collect, process, and analyse the data. The plane takes off in step #2 of this illustration and streams data when it reaches the target.

- Terra8 Agricultural Drone - in Japan, this drone hovers a few meters over a crop and precisely sprays pesticide over 33 acres in one hour. Some farms use 10 or more of these simultaneously.
- Yamaha gas-powered drones can spray 8 liters of pesticide and seed, and use sensors and GPS navigation for precision agriculture.

Drones can also help with water management. Researchers at Texas A&M used a drone to image 7 acres of wheat to find crop disease and avoid putting water on the affected area.

**Computer Science and the Manufacturer**

Farmers also grow the raw ingredients that go into more complex food staples. An example of this is Campbell’s “Chunky Beef with Country Vegetables”. This product contains 27 ingredients including beef, carrots, celery, peas, potatoes, salt, wheat flour, and more. Many of the foods we eat contain multiple ingredients and those foods must be prepared using an assortment of inputs from various parts of the FSC. There is also preparation equipment involved in producing the product like ovens, conveyor belts, labeling machines, weighing and packaging machines, and more. Another major step involves the basic manufacturing infrastructure of the factory building: power and cooling, workers, a cafeteria for the workers, building insurance, and loading docks for the trucks, among others.

Computer science plays a big role in this part of the complex supply chain. Huge amounts of information flow through the manufacturing process to coordinate all the moving parts, such as:

- Ingredients that may arrive JIT or are warehoused somewhere
- Man-machine coordination
- Orders from distributors that need to be priced and sourced to create delivery schedules
- Orders placed with the farms and other sub-contractors, and more

Companies have relied on On-line Transactional Processing (OLTP), Enterprise Resource Planning (ERP) systems, and Enterprise Data Warehouses (EDW) to help operate their organization in these data-rich environments. We are now witnessing the shift from data to
information. Coupled with advances in computer science like cloud computing, big data, platform 3 mobile apps, and others, we soon realize the FSC simply would not work without the technology we often take for granted.

There are many food manufacturing processes that are found in other manufacturing systems. At a high level, a manufacturer uses “ingredients” to “make something” out of them before “shipping” finished products to market. Manufacturers are concerned about the same basic issues such as providing an excellent customer experience, ensuring high quality, efficient operations, inventory control, adherence to rules and regulations, performing required maintenance, and profitability.

Supply chain management (SCM) is critical to keeping a manufacturing plant running at top performance. Demand from the consumer can trigger a wave of demand through the distributor, manufacturer, and the farmer. Rather than just send documents, SCM uses Electronic Data Interchange (EDI) to streamline the information flow so computers can “talk” to other computers in a common language. This helps tighten the vast FSC so information can flow in the chain faster than the goods themselves, making the supply chain more responsive to the consumer. Customer service goes up, data volumes become huge, and the world becomes a smaller place. Here’s an example of a manufacturing process.

**ConAgra and Peanut Butter – A Visit to the Manufacturer**

People all around the world rightfully regard peanuts as a great source of nutrition. As the Wikipedia entry for peanuts shows, they are nutrient rich and a good source of “…niacin, folate, fiber, vitamin E, magnesium and phosphorus. They also are naturally free of trans-fats and sodium, and contain about 25% protein (a higher proportion than in any true nut).”

Peter Pan peanut butter is owned by ConAgra Foods, a $16 billion-dollar company with two dozen food brands and 33,000 employees around the world. Peter Pan makes five types of creamy peanut butter, three types of crunchy, and three all-natural varieties.

Based on point of sale (POS), past demand, marketing campaigns, seasonal forecasts, and other data, ConAgra determines what regions of a country will consume specific varieties and quantities of Peter Pan during a specific time period, such as during the school year. Depending on a farm’s lead time, which could be impacted by conditions such as weather, pests, etc., ConAgra “communicates” with all members of its FSC that it needs large quantities of peanuts.
One factory may use over 100 million pounds of peanuts a year and, during peak demand, produce over 500,000 jars of peanut butter a day\textsuperscript{73}.

Peanuts need five months of warm weather and an annual rainfall of 20 to 39 inches to grow, with a lead time from planting to harvesting of at least five months\textsuperscript{74}. In the U.S., peanut planting is done in April and harvested in the fall\textsuperscript{75}. The nature of a peanut is that it doesn't “store” well, meaning poor storage makes it susceptible to mold; as such the farmer isn't likely going to grow an enormous amount of peanuts unless they know there is market demand.

ConAgra also knows there are multiple suppliers of peanuts in the world with their own market prices. In the U.S. the market price for peanuts as of December 23, 2014 was $427.79 per ton of “Virginia” peanuts (other varieties of peanuts include “Runner”, “Spanish” and “Valencia”)\textsuperscript{76}. A year earlier, the same variety were priced at $468.38 a ton. The market demand is not just for raw peanuts, but for all of the ingredients that peanuts make such as peanut butter, peanut oil, peanut flour, boiled peanuts, dry roasted peanuts, and industrial medicines and textiles. For example, ConAgra’s Wesson cooking oil division markets peanut oil.

The harvested peanuts are loaded on trucks, ship, or rail, and shipped to the plant. ConAgra keeps track of the peanuts en-route to the plant as it does not want its entire order arriving at the same time, and for food safety reasons. Communication is essential to the process ensuring that the right number of trucks with the right tonnage of peanuts arrives at the right time to the right factory.

After unloading at the factory, peanuts are roasted in a 150 foot-long oven. From there, they are blanched (to remove the outer skin), color sorted, grounded twice, homogenized for a smooth texture, and poured into jars moving on a conveyor belt. The modern packaging is a complex process that starts with an injection-molded plastic jar made from super-heated plastic pellets and recycled plastic\textsuperscript{77}. The product is then safety sealed, and given a lid and label. Amazingly, an entire Peter Pan factory is run with just 12 workers\textsuperscript{78}.

The finished peanut butter jars are then automatically placed in cardboard boxes. Depending on the factory, boxes are made on premise or come by truck from a box supplier, the latter having their own supply chain. The boxed product is put on a pallet, given a RFID chip to aid its automated tracking, and shipped to waiting warehouses.

An RFID is a microchip with an antenna that broadcasts a unique code to a receiver without a visual scan. There are three types of RFID – passive, which are powered from the magnetic
field generated by the RFID reader, semi-passive, which has a small battery and is activated by a reader, and active, which has internal power and is always broadcasting\textsuperscript{79}. RFID chips hold 96 data bits, which can be sent up to 20 feet to a receiver/PC that would identify the boxes of food on a pallet\textsuperscript{80}.

Automated manufacturing generates data in every step of the process, and when every machine is digitally networked, it allows the company to have a complete, measurable view of the processes and materials. It goes beyond running the assembly line or growing and increasing market share, to regulatory requirements and consumer demand. A top priority for a food manufacturer is producing safe products, and the data generated during the manufacturing process helps ensure that. Quality controls are integral to the entire process and encompass the entire FSC. In 2007, ConAgra had a major food safety problem when Salmonella was found in their peanut butter. It caused all Peter Pan and rebranded products to be recalled from grocery shelves. ConAgra lost over $150M in monetary damages and additional indirect brand damage. To help avoid similar issues, ConAgra uses International Compliance Information Exchange (iCiX) to automate and manage the collaboration process.

iCiX is used by many industries to help coordinate retailers, manufacturers, distributors, and suppliers throughout their varied supply chains. Using cloud computing, iCiX shares and manages information amongst 20,000 businesses and 70,000 facilities while handling 3 million transactions across 40 countries\textsuperscript{81}. With a worldwide set of producers and suppliers, keeping the flow of information running smoothly is a major computer science problem. As supply chains grow, the increased need for documents and information results in a higher risk of errors and the management of these interdependent processes becomes a big data problem.

The iCiX cloud securely stores, manages, and shares documentation and information between business partners. Through digital rights metadata, the originator specifies which partners can access information. A supplier compliance matrix is established, perhaps with electronically signed and encrypted insurance certificates, product specifications, audit results, and even qualifiers to ensure products are organic or kosher\textsuperscript{82}. Independent inspectors store reports in the iCiX repository so critical data can be immediately acted upon\textsuperscript{83}. Suppliers down to the factory level tie the iCiX data directly into their proprietary systems and a SupplyTrace function creates a map of the food chain. A dashboard shows the originator and suppliers “in” and “out” of compliance, allowing buyers, risk managers, and compliance officers access to relevant data.
Let’s say there was a peanut butter snack cracker recall – a food traceability problem. iCiX’s Rapid Identification of Food Contamination (RIFCOP) tool would broadcast an alert allowing the manufacturer to see who has received it, taken ownership of it, and completed the recall – management by exception. iCiX uses collaborative social media at a product level to achieve a critical goal. The alert may begin with a member’s product inquiry using their iCiX ID and a description of the problem. iCiX responds with a unique tracking ID – i.e. a support ticket. An email is sent to the inquirer, asking for the iCiX IDs for the customers and supplier who had raw ingredients that went into the item or contained in the item, including ITEM#s and LOT#s.

The process continues until all impacted organizations have been notified. This sleuthing approach can prove quite useful when trying to find the source of a problem when many consumers have the same complaint across multiple retailers. RIFCOP can help deduce the common element to the retailers involved – in this example, the highlighted square in the diagram to the right.

Much of the data generated on the assembly line comes from mechanical, electrical, and thermal sensors. A mechanical sensor can gather data on an item’s position or measure its velocity. Electrical sensors can determine voltage, current, and resistance. For example, this sensor measures the presence of liquid with a certain electrical conductivity and can be used in the manufacturing of peanut butter.

Sensors attach to the input module of a programmable logic controller (PLC). A PLC is a small processor with memory, networking, and a special operating system that collects analog and digital measurements and forwards them to a master computer and database. This PLC block diagram shows the I/O modules, CPU, memory, power supply, and non-volatile memory. A sensor would signal the PLC when a jar is full. Output modules could turn on or off a mixing motor, or a peanut butter pump that fills a jar.
A program tells the PLC what to do based on sensor input signals. For example, if a sensor reports a jar is full, the PLC turns off the filling pump, tells the conveyor belt motor to advance, then turns on the pump to fill the next empty jar. “Ladder logic” is a PLC programming language. Assume you had a simple circuit. The battery is wired to the switch and then to the bulb. Closing the switch turns on the bulb. The ladder diagram shows switch S1 and lamp PL1. The horizontal line is like a ladder rung and the side power rails L1 and L2 are the ladder rail. Ladder programs use uniquely addressed graphical symbols for a switch, motor, lamp, etc. For example, with two switches the first would be S1, and the next one S2, allowing the program to uniquely refer to the component being controlled. In this diagram, a mixer motor stirs the liquid in the container when the temperature and pressure sensors reach a preset value, or with a manually operated switch.

Some modern factories are evolving beyond the simple PLC to a PAC. A Programmable Automation Controller (PAC) is essentially a personal computer with a self-contained set of PLCs that could control multiple sets of devices. A PAC can be programmed in languages like C or C++ rather than the graphical ladder logic representation of coils and contacts.

The data collected by PLCs and PACs is sent to a transactional database. While not a typical OLTP, the database nonetheless would capture the details of what the sensors have sent to the controllers with full timestamps, lot numbers, temperatures, and other statistics.

Some of the collected data is used in higher level systems that help run a business, such as an ERP system. An ERP is a very large database that holds a single master record for items like products and employees, with transactional data records for all company activity. It collects, stores, manages, and helps derive business value from sales, invoicing, shipping, inventory management, product planning, marketing, manufacturing, procurement, and accounting daily activities. It can also tie all business functions together through add-on modules such as sales and marketing, finance and accounting, production and materials management, and human resources. Using ERP as a base, additional functionality can be added to encompass areas such as Customer Relationship Management.
Management (CRM), Supply Chain Management (SCM), and Project Lifecycle Management (PLM). In particular, the CRM system can interface with an EDW which targets marketing, fulfillment, sales, and service to the consumer.

The analysis of customer data in an EDW is typically performed by the marketing department and involves a technique called data mining. Data mining tries to find patterns in an EDW that determine relationships like buying patterns which can trigger future customer behaviors. For example, data mining would find a relationship between mothers buying bread, sandwich bags, and peanut butter during school season. From this target audience, ConAgra could increase sales through coupons, or cross-market their Banquet brand frozen fried chicken since they could deduce the mother probably leads a busy lifestyle with young children – a target market for frozen fried chicken. Data inputs into warehouse can come from ConAgra, its vast supply chain, or even Twitter and Facebook data feeds. ConAgra can mine this EDW to determine price fluctuation, package sizing, or many other business problems associated with forecasting demand and increasing sales profitability.

With so much data to process from farmers, manufacturers, distributors, grocers, and consumers, ConAgra moved from traditional server/storage I/O processing to in-memory processing. They selected SAP HANA (High-Performance Analytic Appliance) to handle the high transaction rates and query processing. HANA’s in-memory relational database is column-oriented which is useful for CRM systems. Its design allows it to load structured data from other SAP systems like ERP and other data sources with great speed. By using the server memory instead of a disk, it allows for near real-time analysis applications that interface with HANA. Many HANA appliances use 1-2 TB of RAM, but in 2012, SAP demonstrated a 100 TB HANA system.

**Computer Science and the Distributor**

Further down the FSC, we arrive at the distributor. The farmer has delivered their goods to the manufacturer and the manufacturer in turn has created ready-to-eat or prepared products and shipped them to distributors all over the country. Based on real and forecasted demand,
distributors are buffers that purchase finished products from manufacturers and sell those goods at a markup to retailers like your local grocer. A distributor must have good communication with manufacturers and retailers to have any hope of efficiency since they do not want to pay for and store excess inventory it cannot reasonably sell, nor does it want to miss a market opportunity by not having enough products on hand to sell to the retailers.

Manufacturers like ConAgra need to have enough Peter Pan in the distributor’s warehouse to ensure retailers have enough to sell. If the distributor has too much supply, the product can become stale. If they have too little, a retailer’s “big Peter Pan sale” could empty its shelves, leading to unhappy customers and/or customers who switch brands. The grocer has limited shelf and back room storage space, so it relies on the distributor for JIT delivery. Retailers operate on low profit margins, so if they order too much product from the distributor, they are forced to put it on sale because of the limited space, further eroding profit margins.

Distributors want to control the “bullwhip effect”. In theory, the laws of supply and demand coupled with good communications between supply chain members make a harmonious market. In reality, long food supply chains are subject to cycles of over- and under-production. With a lack of coordination and insight, orders can fluctuate throughout the chain, much like the shape of a bullwhip, hence the name. Market distortions create incorrect forecasts of true demand, triggering higher labor, manufacturing, and transportation costs, larger inventories, longer replenishment times, product shortages, and profitability declines. Supply chain orders get incorrectly amplified as they move away from the consumer, and with overly batched requirements, the bullwhip effect can ruin a supply chain.

For example, let’s say your grocery store plans to have a big two-day peanut butter sale. The store places a large product order with the distributor in advance, who in turn orders more from the manufacturer, which triggers the farmer to bring more peanuts to market. Meanwhile, the sale is so good that consumers empty the grocery shelf of product. The store quickly orders more from the distributor who in turn communicates the market data throughout the chain.

Before too much time elapses however, the consumer becomes disappointed when they cannot buy peanut butter at the sale price. At the same time, the distributor ships a lot of product to that
store. With the distributor’s other customers also running low, they signal the shortage to the manufacturer who starts to ramp up production even with long lead times. Regrettably for the distributor, by the time production has increased, the sale was long over leaving all the members of the FSC with a lot of product but no great market demand. Small variations in order quantities lead to high variations in perceived demand. This economically problematic curve begins to look like a bullwhip\textsuperscript{94}. To minimize the bullwhip effect, ConAgra Foods has worked to streamline processes involving vast amounts of data.

1 The checkout transaction begins when the grocery cashier scans the customer’s loyalty card. As the Universal Product Code (UPC) on the peanut butter’s jar is scanned, the transaction is stored in a POS database. At a logical point, the grocery store’s master database is updated to show a jar of Peter Pan was sold. 2 A store planner is alerted when the shelf stock runs low and notifies a grocery worker to restock the shelf using product in the back room. The POS information may also be relayed to the distributor. 3 When the inventory alert of available stock on hand gets low, it triggers the planner to place (or automatically send) an EDI order to the distributor for cases of peanut butter and any other needed products. 4 The distributor receives the order for cases of Peter Pan and its staff gathers the items and other ordered products. 5 The entire order is palletized and loaded on a specific truck for delivery to that store. 6 As the distributor’s peanut butter inventory reaches a trigger point, a product order is sent to the Peter Pan Company. 7 The truck arrives at the grocer, replenishing store stock and shelves, and adding the inventory to the in-store database. The accounting department is notified that purchases from the distributor are available for sale. Peter Pan prepares an Advanced Shipping Notice (ASN as defined by the EDI), a form of packing slip, when the specified number of cases of peanut butter is being shipped to the distributor.

Decades ago, the grocer relied on postal mail to order products from a distributor. The fax made the process faster, but it was still incredibly difficult to conduct business with delays of days or weeks\textsuperscript{95}. 

---

2015 EMC Proven Professional Knowledge Sharing
Numerous parties could be involved in a simple order with multiple supply chains involved. As we have seen, the grocery store has to issue an order, which is sent to the distributor and possibly the manufacturer. The shipper needs written orders detailing the date the shipment is due and a bank may need to supply credit or process a check as seen in the chart above.

EDI made the same business process a lot faster. It also lowered personnel costs and expenses, improved customer service, reduced errors and out-of-stock events, achieved faster payments, and ensured better control over information. It streamlines the interaction between supply chain members by using an agreed upon structured data standard for electronic transfer of information between each partner's computers. Modern EDI is over 30 years old, with standards that go well beyond our example, allowing companies to exchange purchase orders, invoices, shipping notices, and more. When a store requests a distributor to ship product based on POS information, it is called a “pull”. When the distributor automatically sends a shipment to a retailer based on forecasted demand, it is called a “push”. A “pull” can be exact because it is based on actual demand, while a “push” works well when there is predictable demand.

This sample abbreviated EDI form is used by the Kroger grocery store chain to purchase products from a fictitious distributor. The coding used is in the left column. For example, in the first highlighted row, “G50” means Purchase Order information. The “*” is a separator, so the next field is a “C” which stands for Confirmation. After the next “*” you see the date of January 11, 2005 followed by the Purchase Order number Kroger used for this transaction.
Let’s look at the last three items. G68 stands for an item to be ordered, G69 is a description of the item, and G70 is the code for other detailed information. In this example, Kroger needs 180 boxes of Post brand Shredded Wheat Spoon Size. Kroger orders this cereal by the case at a cost of $31.79 for 12 boxes. The 16.4 ounce box has a UPC code of 004300018062.

While the EDI format is hard to read, it is highly structured and relatively simple for a computer to decode, which is what it is designed for. User-friendly programs use drop-down menus that “speak” EDI. ERP systems also have EDI translations to permit the system to automatically place orders, in this case by generating an EDI 875 document from the system and sending it to another party. The ASN discussed earlier is also an EDI document – EDI 856. There are over one hundred types of EDI documents that can be issued.

The Food Supply Chain Depends on Trucking

With orders whizzing back and forth, transportation logistics are critical. The trucking industry links members of the FSC, tying farms together throughout the country. Farm products and other materials arrive at the manufacturer, and before long, trucks are delivering finished goods to distributors. From the distributors, trucks are dispatched to networks of grocers.

In the U.S., 83% of high-value, time-sensitive freight moves by truck. Trucks move machinery, packages, food, and more. In 2002, the average daily truck traffic looked like the image on the left. By 2035, however, traffic is expected to dramatically increase as shown on the right.

A good amount of the truck traffic will be attributable to JIT delivery, which puts more trucks on the road with smaller payloads. Increased volume means more traffic, greater fuel consumption, longer delivery times, increased truck wear, deteriorating road conditions, longer
passenger car commute times, more accidents, larger inventories to compensate for longer transit time, and added pollution from idling vehicles. Small disruptions can lead to major delays.

The challenges of moving food and other freight by truck need to be addressed by increased efficiencies since building new roads is a major economic burden, environmentally difficult, and politically challenging. Efficiency begins with increased fuel economy of diesel trucks and streamlining operations, which big data can help achieve. With promises of operational efficiency, profit margins can be maintained or improved while reducing fuel consumption.

**Fuel Economy of Diesel Trucks**

In 2011, the U.S. established standards for trucks weighing over 33,000 pounds that would reduce their diesel fuel use 20% by 2018\textsuperscript{101}. Saving 530 million barrels of oil would lower trucking expenses by $50B and reduce or help maintain the cost of food in the United States. Engine and truck makers expect to increase economy from 6.5 mpg to 9.75 mpg through advanced engine and automatic transmission design, increased aerodynamics, weight reduction, reduced tire rolling resistance, hybrid techniques, automatic engine shutdown, and more\textsuperscript{102}. Biodiesel could also decrease some exhaust emissions and reduce national reliance on foreign energy sources\textsuperscript{103}.

**Streamlining Operations**

Creating an efficient trucking industry means more than efficient trucks. It means using technology to help 2 million tractor trailer drivers who log 140 billion miles a year, including sensor data on truck performance, providing routes for lowest cost/shortest time for delivery addresses based on truck loading patterns, using GPS best-route redirection based on road or traffic conditions, and other real-time analysis\textsuperscript{104}. This data can also predict maintenance requirements, inform mechanics about problems so parts can be gathered, and determine if servicing can be done during scheduled downtime or with minimal disruption.

Data from the truck also helps the logistical system track food in-transit. It can be used to predict JIT arrivals and departures, reduce road congestion and air pollution, help insure driver safety, aid in making last minute changes to optimize deliveries, and allow alternate transportation to optimize cost, time, and customer experience, all to create a more efficient supply chain. For example, with bad weather, it could make sense to drive slower and let a storm pass than to drive faster and be affected by it. Truck data allows the driver to make informed decisions. Collecting and processing this data in real-time is a big data problem. A driver must take breaks, so idle time, truck braking habits, engine temperature, and more are areas for improved
efficiency. The North American Council for Freight Efficiency executive director said in 2012 that "Between the worst driver and the best, the difference in fuel economy can reach 25 percent."  

Logistical efficiency comes when the right size truck is used to move a given amount of freight while spending less time waiting between loads. One study showed that 29% of trucks on the road were empty with 58% of semi-trucks half-empty or worse, leading to fewer tons of freight per gallon of fuel, more trucks on the road than needed, and unnecessary pollution. Engine idling allows the truck to maintain tractor refrigeration of perishable food. However, idling a 500 hp engine to power the air conditioner doesn’t make economic or environmental sense, so more truck stops are being equipped with electrical tethers. Some transportation companies like Followmont Transport in the State of Queensland, Australia, and U.S. Xpress headquartered in Chattanooga, Tennessee have made major efficiency gains through computer science.

Followmont Transport

Followmont Transport is a 30+ year old Queensland freight company with 600 trucks and 19 depots. They make 130,000 deliveries of mail, bread, magazines, pharmaceuticals, produce, and general cargo a month, making it the largest independent hauler in that part of Australia. As a medium-sized business, they made a sizable IT investment over the years, yet daily and weekly reports often arrived too late to help make business decisions. In 2008, operations ran on 15 physical servers with direct-attached storage, no virtualization, and no disaster recovery capability.

Paul Smith became their CIO in April 2011, and he overhauled IT operations to help improve efficiency. They now have two data centers in case of a disaster with modern firewalls and intrusion prevention hardware. Each data center uses the latest Dell PowerEdge virtualized servers and a 180 TB 3-tier Dell Compellent storage system. They also upgraded their freight management system that in addition handled their CRM and billing, with a cloud-based supply chain and logistics Software-as-a-Service system called CargoWise One, that integrates, automates, and communicates with the supply chain.

Much of the Followmont efficiency gains are attributable to SAS’s Visual Analytics business intelligence (BI) system that helps them make rapid decisions from visualized real-time data. The latest information, along with a 12-month historical view of revenue, tonnage, and other data, is available to drivers in their cabs and depot managers through an iPad. If they find an issue, they can address it immediately with the customer. They have reduced underperforming runs and raised customer satisfaction. Paul Smith says “Now our sales and depot managers..."
can see this information quickly and take action to improve the freight mix to make it more efficient.” Customer retention has also increased through the use of discounts to visually valuable customers. The systems they put in place have also increased truck utilization by being able to fluidly mix bulk and carton freight.

Customers now access their information through a web portal for full reports, invoicing, summary billing, and shipment details. The system helps Followmont control expenses by analyzing wages and costs to deliver freight to different locales. Additionally, vehicle maintenance gets more attention by allowing mechanics to provide the proper preventative attention and keep the truck fleet modernized. Smith says, “If you’ve got leased trucks you want them to have the same kilometers at the end of the lease. You don’t want to have one with 3 million [kilometers] and one with 300,000. So we can rotate [our] fleet around depending on usage.”

**U.S. Xpress**

Followmont’s fleet of 600 trucks is nonetheless small compared to U.S. Xpress which is America’s 2nd largest privately owned carrier with over 16 times that number. Their fleet hauls food, beverage, and grocery freight throughout the U.S., Canada, and Mexico. Years ago, they felt there was the potential to save millions of dollars on fuel costs if they could find out how much time a truck spent not actively hauling freight – i.e. an idle truck spends money while making none. If they could measure it, then they could manage it.

As discussed earlier, idling a truck can be warranted to keep freight at a certain temperature in an air conditioned trailer or keep the cab at 70°F to allow the driver to sleep during scheduled break time. In the case of creature comfort, Xpress wants the driver to raise the temperature to 78°F-79°F after two hours to lower fuel consumption. It turns out that keeping a tractor trailer truck’s engine idling cost about $2 an hour depending on the cost of diesel fuel. Multiply that by 10,000 trucks on the road, with some idling for hours a day, and the potential for savings is enormous. Saving costs help keep Xpress competitive; the key for them is getting the data.

The company started collecting the data in 2008 when they installed DriverTech devices in cabs. It provides on-board intelligence through data sensors, and is supported by a centralized fleet management system that receives over 958 data elements from
each truck every 15 minutes\textsuperscript{118,119}. This system creates over 9 million real-time data points covering how fast the truck is going, diesel usage, tire pressure on up to 18 wheels, brake conditions and the number of hard braking events, engine status, GPS coordinates, and more.

Trying to derive business value from this data was not easy. Back in 2008, U.S. Xpress had over 130 different databases that lacked integration, 90 mainframe screens, and poor data quality (they had 178 different ways of spelling “Walmart”). Their data came from AS/400 LPARs (IBM System i) running their ERP system and Intel servers. There was no centralized database. “Efficiency” meant getting reliable data in weeks\textsuperscript{120}.

The 958 data points from 10,000 trucks are sent every 15 minutes by cellular Wi-Fi to their mobile carrier, who forwards it over the Internet to Xpress systems in Tennessee. Some of the data deals with truck operations and the rest is geospatial letting Xpress know if a truck was on a highway, at a depot for loading, or being serviced\textsuperscript{121}. To tackle this massive data ingestion problem, they built a Hortonworks front-end to stream the real-time telemetry into a Hadoop cluster\textsuperscript{122}. Sqoop handles bulk transfers between relational databases and Hadoop, and Flume collects, aggregates, and moves streaming data from different sources to a central Hadoop Distributed File System (HDFS)\textsuperscript{123,124}.

Once they addressed the data quality issue to obtain a “single version of the truth”, the operational truck data is passed into a Teradata EDW. For this step they leveraged Informatica tools including Data Quality, Identity Match Option, and Data Explore\textsuperscript{125}. Informatica’s PowerExchange is used to interface with the IBM i systems\textsuperscript{126}. With transactional data still residing in 130 databases, data had to be brought together to make an end-to-end Hadoop real-time data feed by Pig and Hive/SQL.
end information flow. The CTO responsible for this IT overhaul at U.S. Xpress is Tim Leonard, whose background included working at Dell Computer as an EDW Strategist and BI Lead\textsuperscript{127}.

This EDW also supports Xpress’s daily operations. For example, a new customer’s order is assigned an ID and is updated from the operational data store (ODS) using OLTP. As the order ages, it becomes historical EDW content. Xpress processed billions of records to save over $6 million a year – a testament to problem solving with big data and real-time BI\textsuperscript{128}. They found drivers who overly idled trucks and alerted their managers. An historical search found one truck had idled its engine for 7 days\textsuperscript{129} Idling was cut 30% in 4½ months of system use, saving $20M in fuel costs the first year\textsuperscript{130}. The data also helped improve truck upkeep to save another $1.2M a year, monitor mandated maximum hours behind the wheel, and detect unsafe driving habits. Eventually Xpress integrated a refreshed CRM system from Microsoft which, when coupled with fuel consumption awareness, allowed fleet managers to gain better insight into their trucks’ operations and improve customer relations, all through their iPad\textsuperscript{131}.

**Computer Science in the Grocery Store**

A supermarket is a step up in size from a grocery store, stocking a full line of meat, produce, canned goods, dairy, household items, and more. As of 2013 there were 37,000 in the U.S. employing over 1% of the country. Annual supermarket sales were $620B or almost 4% of the Gross Domestic Product. Average stores exceed an acre in size with nearly 44,000 items and sales of $318,462 a week\textsuperscript{132,133}. Families with children spend over $8,000 a year on groceries\textsuperscript{134}.

**The Barcode “Miracle”**

You’re at the supermarket and just finished shopping. Your items are moving down the conveyor belt where the cashier scans each item at the POS terminal. The POS display shows what is purchased; meanwhile, the system totals the amount and applies sales tax. Behind the scenes, a record of the transaction is sent to the inventory control system, and customer information to the CRM database. If inventory runs low, the system issues orders for additional stock. When you swipe your credit or debit card through the card reader, the information is transmitted to the card issuer for approval. After signing the authorization screen, you are handed a receipt. With the items bagged, the cashier is ready for the next shopper.

The modern shopping experience would not exist if the barcode was not invented. On October 7, 1952, Bernard Silver and Norman Joseph Woodland received US Patent #2,612,994 for a
barcode scheme to represent the UPC\textsuperscript{135}. Retail sales were forever transformed when the barcode on a pack of chewing gum was scanned on June 26, 1974 at an Ohio supermarket.

When the barcode is scanned, the information appears as a data string to a POS system. This data goes into a transaction processing system (TPS), sometimes called OLTP.

**How does a barcode work?**

Just like Morse code encodes text with dots and dashes, such as SOS \textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet\textbullet, a barcode encodes text with thin and thick bars. A barcode scanner uses a photodiode to capture the intensity of reflected light off the white spaces in order to measure the width of the bars and spaces (since black absorbs light). Depending on the length of the barcode and its dimensions, it can contain information about the manufacturer, flavor, size, price, product name or in the case of U.S. grocery barcodes, a manufacturer and a product’s code number.

The Uniform Code Council in Dayton, Ohio assigns a unique 12-digit code to each U.S. retail product. In the U.K., retailers use a 13-digit “EAN-13” code which identifies the product’s name, its country of origin, and the manufacturer’s details.

The UPC has a left and right segment allowing the code to be read in any direction. It begins with a “101” start and stop character. The left has 6 numbers (012345) to represent the “Manufacturer ID #” and are selected from “Character Set A”. The center guard is “01010”, and the right 5 numbers (54678) are the “Product Code ID #” and are from “Character Set C”.

There are 12 “bar” characters encoded in UPC-A, representing 11 data digits and a 12th self-check digit for error detection\textsuperscript{136}.

To compute the “check”, multiply the odd location digits (1,3,5,6,7,11) by 3 and sum them (0+6+12+15+18+24=75). Add the even numbered location digits (1+3+5+4+7=20) and add them to the odd (75+20=95). Perform modulo 10 math on the sum (95 mod 10 = 5) and subtract this from 10 (10-5=5) which should equal the check digit in green above. This determines if the photodiode correctly scanned the barcode. This recorded wave pattern shows the “high” parts are the absorbed “1”, and the “low” the reflected “0”. The code,
with manufacturer’s ID of 012345 and the product’s code ID of 54678 is then passed to the POS machine. In this scan of a chicken noodle soup can, the barcode reader found a manufacturer’s code of 051000 and a product code of 01251. You can look up a UPC code at //upcdatabase.org/ or with a smart phone app at //scan.me.

Our example used one of 30 linear barcodes\(^{137}\). There are also 24 matrix codes, each capable of conveying a lot of data. The one on the far right is a Quick Response Code (QR code) which encodes 4,296 alphanumerics or 7,089 numbers\(^{138}\). The QR is very popular on advertisements.

**How does the Point of Sale Terminal work?**

A POS is essentially a PC with peripherals such as a cash drawer, customer and cashier display, receipt and loyalty coupon printer, check imprinter, and a produce scale. Data entry devices are attached that make the operation flow with speed, efficiency, and precision, such as a scanner and credit card reader. These peripherals typically connect with RS-232 DB9F or USB cables. The PC is a modern computer with a dual core processor, RAM memory, dual display graphics controller, SATA hard drive, and an array of ports to plug peripherals into. They run Microsoft Windows, DOS, Linux, or a proprietary operating system. Inside you would find a ruggedized motherboard.

**What Does the Data Flow Look Like?**

A typical customer transaction begins with the customer’s loyalty card. It has the customer’s assigned ID number which is reproduced as a barcode. When they requested the card, demographic data was obtained on their marital salutation (Mr., Mrs. or Ms.), postal and email address, age, and other data. The information was stored in a CRM database on a backoffice server. As the customer purchases groceries, this database is updated with their brand preferences and the date they made those choices. The data can then be mined to determine when to print a POS coupon encouraging them to buy more Cheerios cereal, for example. Additional 3rd party data may augment this database, such as the likely number of children in the family, their sex, and ages (although it is possible to deduce information based on the items frequently purchased – i.e. Fruit Loops cereal, lolly pops, diapers, etc.)
As each item is scanned, a time/date stamped record reflecting the item’s UPC code is recorded and the local price database is queried for the item’s current price, which should also match the price the customer saw on the item’s shelf. The price database knows the UPC represents a 15oz box of Cheerios and costs $3.95. Even products made “in house”, such as a fresh, store-baked apple pie or deli-cut roast beef have barcodes, but these refer to in-house stock keeping units (SKU – the UPC code is a subset of the SKU) and not a manufacturer’s code. This physical flow diagram shows the interaction within the POS during the lifecycle of a completed customer transaction.

At the end of the transaction, the clerk scans any coupons and presses the “Total” key. The system computes any sales tax and displays the total amount due. Customers paying with credit or debit cards have a magnetic stripe or chip containing their ID, and when read by the card reader a data record is sent to the POS computer. The POS transmits it and the amount due to the processing server or credit card processing system. The CRM database may also get an update from the credit card vendor. Finally, a detailed receipt and coupons for future purchases are printed. The completed transaction is sent from the POS computer to the backoffice server. The backoffice receives every POS transaction from every store and batch processes them. In some stores, the POS terminal maintains every transaction detail until polled for its full tally of transactions, perhaps at night when the supermarket closes.

As the backoffice system updates its master and inventory database with units sold, other reports are created such as an ASN to instruct the distributor to deliver more cases of Cheerios to that store, or to inform the store restocking manager to put more Cheerios on the display shelf. As the distributor delivers a truckload of items, they are added to the store’s inventory.

If a supermarket company has many stores, data aggregation across them is likely done as a batch function. Some supermarket chains feed all their individual store data into into a single analytical system – i.e. an EDW, management information system (MIS), decision support
system (DSS), BI, executive information system (EIS), or expert system (ES). The flow is depicted in this diagram.139

Data Warehousing and Big Data

As we said earlier, supermarkets have very low profit margins. They are always battling competitors like drug and convenience stores, big box retailers (e.g. Walmart), and e-retailers that apply pricing pressure to the FSC and threaten to further erode the supermarket market share. Some supermarkets counter by selling higher margin store brands, prepared foods, and items like greeting cards, electronics, and magazines that are found at their competitors.

Supermarkets also try to create a competitive beachhead through the use of loyalty programs that promise to help retain their existing customer base. With the promise of added discounts and an improved shopping experience, the customer’s data is also used to help arrange the store’s layout to increase the profitability and quantity of items in their shopping cart.

SAS’s grocery Revenue Optimization Suite “…enables grocers to achieve stronger financial results – increased sales, profitability and gross margins – by proactively planning, optimizing and implementing both regular and promotional pricing based on a thorough understanding of consumers, demand and the market.”140 They apply data modeling and forecasting analytics to understand customer clusters and portray demand for store merchandise. For example, the CRM system contains demographic data such as a customer’s propensity to buy certain foods at holiday time. The dynamic models change as shopping patterns change. They can leverage other approaches to big data retail operations by combining internal sales data with social media data, such as Facebook “likes”, to help forecast demand. Some models are obvious, like increasing weekday store traffic by discounting roasted chicken 50% from 4-7 P.M. to attract busy parents. Once there, they are likely to buy other higher margin items.141

Catalina Marketing is an in-store behavior marketing company focused on customer retention and increased margins. When their EDW analysis of your purchases shows you would likely buy a product they promote, you receive their POS coupon. Rather than send everyone a coupon, they focus direct marketing costs on segments showing a propensity towards buying that product. For example, an analysis may identify you through your shopper card as someone who
only buys Oreo cookies when they are on sale, and has not done so in a while. As a result their analytics print you a milk and Oreo coupon for your next visit.

**How do you analyze all that data?**

Data is coming in from many places – internal transactions, sensor streams, weather patterns, social media, and many other sources. Some of these sources have a direct business impact while others are just “noise” and are of no value – e.g. the direction an Argentinian bird is flying will have no impact on your decision to buy cereal. With the goal to gain useful knowledge from data, how do you make sense of it all and decide what is relevant or not?

Data science, another computer science practice, focuses on formulating context-relevant questions and hypotheses that yield actionable outcomes. A data scientist’s job is to identify and transform data sources to produce those outcomes, or demonstrate via statistical evidence that desirable patterns exist. Much of this work involves data modeling and the “art” of inference. Data scientists also have a feel for whether an EDW or Hadoop will best solve the business’ problems. Both designs can handle large data volumes, so expertise is needed to identify workloads that run best on a certain platform or architecture, best suited to a particular company, or applicable to a use case.

At a high level, an EDW excels when data is structured in rows and columns, or has defined field lengths. It also is very good at joining subject areas and offers high performance when BI tools are needed.

Hadoop is superb when the data is raw or lacks structure and is complex. It also allows for the use of numerous programming tools.

There are also use cases where EDW and Hadoop work together, and some features from Hadoop are finding their way into EDW solutions. This chart was jointly co-authored by Teradata and Cloudera, leading suppliers in the EDW and big data movement, respectively.

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Data Warehouse</th>
<th>Big Data Hadoop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low latency, interactive reports, and OLAP</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>ANSI 2003 SQL compliance is required</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Preprocessing or exploration of raw unstructured data</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Online archives alternative to tape</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>High-quality cleansed and consistent data</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>100s to 1000s of concurrent users</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Discover unknown relationships in the data</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Parallel complex process logic</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>CPU intense analysis</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>System, users, and data governance</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Many flexible programming languages running in parallel</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Unrestricted, ungoverned sand box explorations</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Analysis of provisional data</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Extensive security and regulatory compliance</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Real time data loading and 1 second tactical queries</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

**Data Warehouse Basics**

Data warehousing can trace its antecedents to traditional information technology (IT) of the 1980’s. At that time, Teradata was the major vendor in this space with a parallel relational database management system (RDBMS) that ran on massively parallel processing (MPP) servers with the workload distributed amongst the nodes. Each server runs independently with a...
"shared nothing" architecture. An EDW does not need a parallel architecture, but parallelism greatly increases its performance and is important when data is measured in terabytes.

Data warehousing has two founding fathers. They are Bill Inmon, whose bestselling 1992 book, “Building the Data Warehouse”, sold over 500,000 copies, and Ralph Kimball, whose 1996 book “The Data Warehouse Toolkit” sold over 375,000. Yet their approaches to building an EDW are vastly different. Inmon advocates a top-down design where a normalized data model for the warehouse is designed first. Dimensional data marts (DM) that tie directly to specific business processes or departments come next. They are designed from the EDW with a departmental or specific view of data - a grand-plan approach. Kimball uses a bottom-up design where the DMs that produce the businesses’ reports and analysis are built first, and are brought into an EDW later on. This chart summarizes the pros and cons of each approach.

<table>
<thead>
<tr>
<th>Top-Down Inmon Design</th>
<th>Bottom-Up Kimball Design</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pros</strong></td>
<td><strong>Cons</strong></td>
</tr>
<tr>
<td>Corporate-wide endeavor</td>
<td>Longer project</td>
</tr>
<tr>
<td>Designed from start rather than evolved</td>
<td>High risk, highly complex</td>
</tr>
<tr>
<td>Single data repository</td>
<td>Large teams</td>
</tr>
<tr>
<td>Well defined</td>
<td>Expensive</td>
</tr>
<tr>
<td>Easier to enhance and enrich</td>
<td>ROI from results can take a while</td>
</tr>
</tbody>
</table>

Whichever approach is used, data must be extracted from various sources, transformed in a uniform representation, and loaded into a database. This step is abbreviated ETL. Corporate data might come from spreadsheets, Word documents, or PDFs, and be located in files, small databases, vendor transmissions, and more. Even within their context, data can take on different meanings, formats, and representation. ETL is a resource-intensive, continuous process to add new data from various sources into the EDW or DM that represent a “single version of the truth” – i.e. data values with unique definitions.

One major part of ETL is data standardization. As we saw earlier, U.S. Xpress had 170 ways of spelling their largest customer, Walmart. Achieving a common data view when merging sources
can be a daunting task. Just look at the
problem that names and addresses cause the
U.S. Post Office. When optically scanning
postal mail, they recognize seven spellings for
Avenue. The problem is further complicated
when key words and abbreviations must be determined contextually. For example, a brute force
transformation of this fictional address leads to a lot of confusion. Is “MS” a salutation, mail-stop,
or state abbreviation? Is “ST” part of a
person’s middle or last name? Is it part of a
church, a town, or street name? As a
supermarket collects customer data it needs to standardize the information or face duplication
or even worse, the possibility of drawing incorrect conclusions increases. For example, does
“Ms.” refer to Jill St. John’s salutation or where she resides (Mississippi)?

The complexity of a supermarket’s data could involve combining dozens of Oracle tables, parts
of many small SQL tables, flat files, and spreadsheets where the first and last names and
birthdates need reordering, all into a common format. This is especially true when supermarket
chains have grown through acquisitions where the original organization had their own data
definitions. Companies specializing in this critical step are Informatica, IBM, SAP, and Oracle.\(^{145}\)

A supermarket’s EDW architecture is the
incorporation of many
operational systems
including CRM (loyalty
cards), TPS/OLTP (POS
transactions), ERP
(logistics between
manufacturer, distributor,
and supermarket), legacy
data, and external 3rd
d party data. When brought together into an EDW, the size can be
measured into 100’s of terabytes or more. In this data flow, source systems\(\) are transformed
instantly or in batch cycles, and then loaded into the warehouse\(\). Specific data is put into
DMs\(\), and analyzed\(\). Some environments use the ODS\(\) as the focal point for real-time
transformation prior to warehouse integration.
Many tools are used to measure, report, analyze, predict, and forecast a company’s situation and likely outcomes. BI is part of a $14B market of tools featuring online analytical processing (OLAP), ad hoc query abilities, and basic tools like Excel. Some longstanding leaders in the space are Microsoft, SAS, SAP, IBM, MicroStrategy, and Information Builders. When BI tools are used they rarely operate on the EDW itself, but rather against routinely extracted data placed into DMs (Inmon’s approach).

Data can be used for a host of activities, such as improving store sales (as we saw in the SAS example) or in direct advertising such as around holiday time. For example, if Alice orders a large turkey then she is likely planning a large party, so coupons and promotions for accessories could be emailed or texted to her smartphone: “Alice, with the holidays just around the corner, we want you to know there is 30% off sale on napkins, stuffing, roasting pans, and frozen appetizers.” In-store directions displayed on Alice’s phone could show her where items are located in the store and can even access her grocery list prepared at home, all to enhance her shopping experience. The visualized data can also be used by store management to adjust inventory and in-stock positions, which can be very useful with perishable food like meat, dairy and fresh flowers.

The data in the EDW can use different designs, from Inmon’s relational 3rd normal form model (think of removal of redundant data and establishing table relationships) to Kimball’s dimension “star” and “snowflake” schema. Many EDWs use Kimball’s schemas where dimension tables surround a central fact table and are shaped like a star. The Kimball version is easier for business users and tends to perform better for analytic queries.

The Kimball star schema below has a “fact” table and four “dimension” tables – geography, customer, product, and time. A fact table contains numeric facts or measurements such as quantity and currency, and pointers (foreign keys) to dimension tables. Fact tables can have billions of rows, all accessible by key dimensions such as employee_code, product_code, etc. A dimension table contains descriptions of the facts. When a join is needed to respond to a query, it means that a fact and dimension tables are “glued” together using the fact table’s primary key,
as well as the foreign key of one or more dimension tables. Indexes speed up data access and are usually associated with a table. A view is a preset query against tables in a schema.

In this supermarket star schema, POS transactions are captured in the center fact table\(^{148}\). The supporting dimension tables include date data, information about the store, a full description of the product purchased, and any promotion used to sell the product.

**Hadoop Basics**

Doug Cutting, who is credited with inventing Hadoop in 2005, gave it the name of his son’s stuffed yellow elephant. Using a Google framework designed to rapidly ingest web data, answer a query, and search data in parallel, he developed an open source project to handle billions of searches and index millions of web pages while working at Yahoo! Hadoop has become synonymous with big data, and from that standpoint, Hadoop and EDWs have a lot in common\(^{149}\).

Both Hadoop and EDWs scale to huge volumes of data, run in parallel (although they can run on single nodes), and have shared-nothing architectures. What sets a big data architecture
apart from an EDW is that it handles huge structured and unstructured data feeds such as clicks, social media, and more. With unstructured data, quotes and commas may not delineate fields. Big data is also known for volume, but volume alone is insufficient to make it big data. It must also have a high data velocity and a variety of formats.

Big data does not require Hadoop but it makes the job easier, faster, and cost less than an EDW. Hadoop contains HDFS, MapReduce, and other components, and in general can process data far faster than serial EDWs. It uses commodity servers with internal drives to break apart batches of structured and unstructured data as well as the raw compute power of commodity processors running in parallel. HDFS spreads chunks of data across the servers without a schema. MapReduce is both a Mapper doing most of the data crunching or extraction, and a Reducer combining or transforming solution sets into a single result. Here are some of the components of the Hadoop framework and what they do.

- Hadoop Distributed File System (HDFS) - fault-tolerant design used on commodity hardware, data spread across all nodes
- Hive - data warehouse system built on top of Hadoop for analyzing large datasets
- Hbase - column-oriented database for random, real-time R/W access
- MapReduce - programming model for processing large clusters of commodity hardware, leveraging parallel processing power of distributed file system with large data sets
- Flume - distributed services that collects data from different sources
- Sqoop - import RDBMS data into Hadoop and vice versa
- Hilo - move data between database and Hadoop
- Chukwa - display, monitor and analyze results of the large collection of logs

Some customers use Hadoop-EDW hybrids because it is difficult to fit unstructured data into a rigid schema, while Hadoop doesn’t use such a format. MapReduce groups, aggregates, and interprets data only when analytical MapReduce functions are run against the data. This is similar to indexing the data in a schema-based EDW to achieve efficient record access. Hadoop does not need indexes. In effect, a dynamic schema is created when the data is read. It is not unusual for Hadoop to feed or access data in the EDW. There are many designs.
The power of Hadoop promises to invigorate the customer’s shopping experience. Some ideas leverage their smartphone, loyalty card, and current location in the grocery. For example, as you pass by the salad dressing shelf, you could be reminded that it has been a while since you bought French dressing, and if you did so today, a 20% off coupon would be applied to your loyalty card for use at the POS that same day.

This same promotional idea can be enhanced when intelligent grocery carts become available. As discussed for many years, a Wi-Fi cart with a display and scanner will track purchases\textsuperscript{153}. With the shopper’s permission (so as to not get annoying), the device can cross-promote items of interest to them or an analysis could trigger a product promotion, achieve higher store revenue, and obtain higher margins. Alice could be offered a private sale on croutons to go with the bag of pre-mixed lettuce she just picked up in the produce aisle. There are obvious food patterns throughout the store, such as when a meat selection is made and Alice already has frozen vegetables in her cart, a promotion for baking potatoes can be offered. The supermarket can also get additional funding from manufacturers by promoting their new product on-the-spot. As Alice walks through the cookie aisle, an offer could appear for a new brand that would appeal to her children yet have lower calories and less sugar.

CRM datasets can be mined using MapReduce to find many-to-many mapping associations. For example, it can help a supermarket identify patterns in buying habits across all shoppers, not just Alice. By identifying items she purchased together, multiplied by thousands of other shopper transactions, dependencies can be found. This Market Basket Analysis (MBA) can address questions like “Which pair of items do people often buy?”\textsuperscript{154} Assume there is a very large sample of supermarket transactions (T#):

\begin{itemize}
  \item T1: cracker, ice cream, beer
  \item T2: chicken, pizza, coke, bread
  \item T3: baguette, soda, herring, cracker, beer
  \item T4: bourbon, coke, turkey
  \item T5: sardines, beer, chicken, coke
  \item T6: apples, peppers, avocado, steak
  \item T7: sardines, apples, peppers, avocado, steak
\end{itemize}

An EDW is going to have a hard time producing an answer, but for MapReduce, the process is simple. It would first distribute the pairings to Map nodes in its system before pairing them:

\begin{itemize}
  \item T1: cracker, ice cream, beer
  \item T2: chicken, pizza, coke, bread
  \item T3: baguette, soda, herring, cracker, beer
  \item T4: bourbon, coke, turkey
  \item T5: sardines, beer, chicken, coke
  \item T6: apples, peppers, avocado, steak
  \item T7: sardines, apples, peppers, avocado, steak
\end{itemize}

The output is then based on (key, value) (pair of items, # of occurrences):

\begin{itemize}
  \item ((cracker, ice cream), 1)
  \item ((beer, cracker), 1)
  \item ((beer, ice cream), 1)
  \item ((chicken, pizza), 1)
  \item ((chicken, beer), 1)
  \item ((chicken, bread), 1)
\end{itemize}
The data aggregation/combination step then gives the results (key, value): (pair of items, # of occurrences):

\[
\{(\text{coke}, \text{pizza}), 1\} \\
\{(\text{bread}, \text{pizza}), 1\} \\
\{(\text{coke}, \text{bread}), 1\} \\
\ldots
\]

Based on this information, items can be physically placed closer to others in the store, promotions can be offered, and customers get a better shopping experience.

This is not science fiction but easily doable technology based on insights into data and the identification of customer shopping patterns – something that Hadoop excels at. Years ago, using an EDW, Walmart increased sales of beer and diapers by 20% to men on Friday nights by colocating these items after finding a shopping pattern. Walmart is also known to feed weather data into their EDW giving them 48 hours to position umbrellas to the front of a store when rain was forecasted. Sales of Pop-Tarts and beer increased dramatically before a hurricane was going to hit, perhaps by people originally setting out to buy a flashlight and batteries. Hadoop makes it easier and cheaper to ask questions of a supermarket's data and have it tell them something they didn't already know.

**Conclusion**

Our goal in *How Computer Science Helps Feed the World* was to demonstrate how computer science is making a tremendous impact in feeding an increasingly growing global population, enabling greater productivity and efficiency for farmers and suppliers while adding value to customers experiences around the world. Geotagging, computer-driven tractors, big data, and more are allowing farmers to maximize the value they get out of their crops, while robotic (or automated) milking has provided a much-needed boost to global milk production. Distributors like ConAgra now rely extensively on computer science to reduce traditional inefficiencies in the FSC, benefiting their consumers while increasing their profits. Trucking companies like Followmont and U.S. Xpress are now using similar technology to get a similar “leg up” on their competitors, while consumers are getting a whole new, better shopping experience from increased interaction with data-driven grocery stores and supermarkets.
It is on the back of computer science-based innovation that the demands of feeding the world will be met and mastered. What can we expect computer science to deliver in the years ahead? As documented by Farm Journal Media’s “Farm of the Future” website, the best may be yet to come. Farmers will be able to directly communicate with consumers in advertising the value of their particular product, using a live-feed of what’s going on at the farm\textsuperscript{155}. Robotic milking will continue to gain in popularity for dairy farmers worldwide, while one expert predicts that cattle farmers will be able to use their smartphones to check on and diagnose the health of their herds\textsuperscript{156}. Even now, the technology-driven economic engine known as Silicon Valley is turning its attention towards high-tech farming in the Salinas Valley, known as the “Salad Bowl of the World”, to pioneer new breakthroughs in agricultural technology\textsuperscript{157}.

The rest of the FSC will continue to benefit, too. The website Overstock.com has launched a beta program to create an online Farmer’s Market, able to ship fresh produce to consumers across the San Francisco Bay area with just the click of a button, with future programs planned for major metropolitan areas such as Houston, Atlanta, and Raleigh-Durham\textsuperscript{158}. 3-D printing may allow companies to produce the same farm equipment at a vastly reduced cost, meaning they can lower prices while increasing their overall bottom line\textsuperscript{159}. The same technology may also allow farmers to make their own equipment without exorbitant time, labor, and monetary costs, greatly increasing the odds that small business farmers will continue to thrive in an increasingly global market\textsuperscript{160}. This too would help lower the costs traditionally passed on to consumers in order to make a profit.

Agriculture has come a long way since the first farmers sowed their crops and domesticated their animals. The goal of our ancestors remains the same for today’s global agricultural economy, however – how can we feed our civilization, and how can we use technology to help us? Today computer science is answering that second question, and we can expect that answer to remain the same for decades to come.
Appendix – List of Abbreviations

AMS - Automated milking System
ANS - Advanced shipping notice
APS - Area positioning system
BI - Business intelligence
CRM - Customer relationship management
CTF - Controlled traffic farming
DM - Data marts
DSS - Decision support system
EC2 - Elastic compute cloud
EDI - Electronic data interchange
EMR - Elastic MapReduce
EDW - Enterprise data warehouse
EIS - Executive information system
ERP - Enterprise resource planning
ES - Expert system
ETL - Extract, transform, and load
FSC - Food supply chain
GLONASS - Global navigation satellite system
GPS - Global positioning system
HANA - High-performance analytic appliance
HDFS - Hadoop Distributed File System
iCiX - International compliance information exchange
IT - Information technology
JIT - Just-in-time
LNM - Lifetime net merit
MBA - Market Basket Analysis
MIS - Management information system
MPP - Massively parallel processing
ODS - Operational data store
OLAP - Online analytical processing
OLTP - On-line transactional processing
PC – Personal computer
PAC - Programmable automation controller
PLC - Programmable logic controller
PLM - Project lifecycle management
POS - Point of sale
QRC - Quick response code
RFID - Radio frequency identification tag
RIFCOP - Rapid identification of food contamination
RDBMS - Relational database management system
RTK - Real time kinematic
S3 - Simple storage service
SaaS - Software-as-a-service
SCC - Somatic cell count
SCM - Supply chain management
SKU - Stock keeping units
TPS - Transaction processing system
UPC - Universal product code
...such as “wholesale distribution business.” In this paper, the term “distributor” refers to both distributor and wholesaler.

Distribution refers to the activities of a business that acts as a middleman between manufacturers or wholesalers and retailers who sell the products to consumers. However, the difference between wholesale and distribution is sometimes blurred, with commentators using the terms interchangeably or combining them in a phrase such as “wholesale distribution business.” In this paper, the term “distributor” refers to both distributor and wholesaler.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.