Best Fit Engineering for the Analytics of Things

Dan Graham, Teradata
Dave Shuman, Cloudera
September 2017

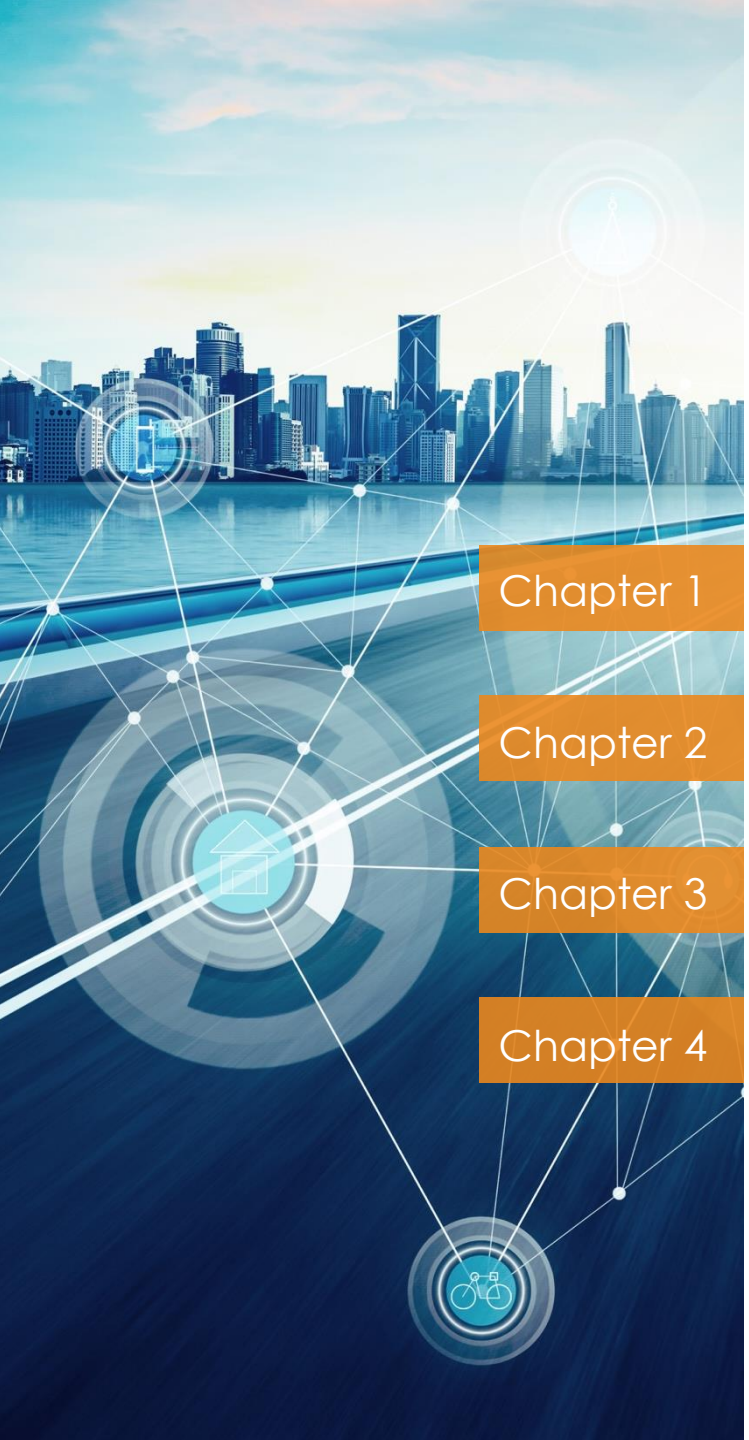


Table of Contents

cloudera TERADATA

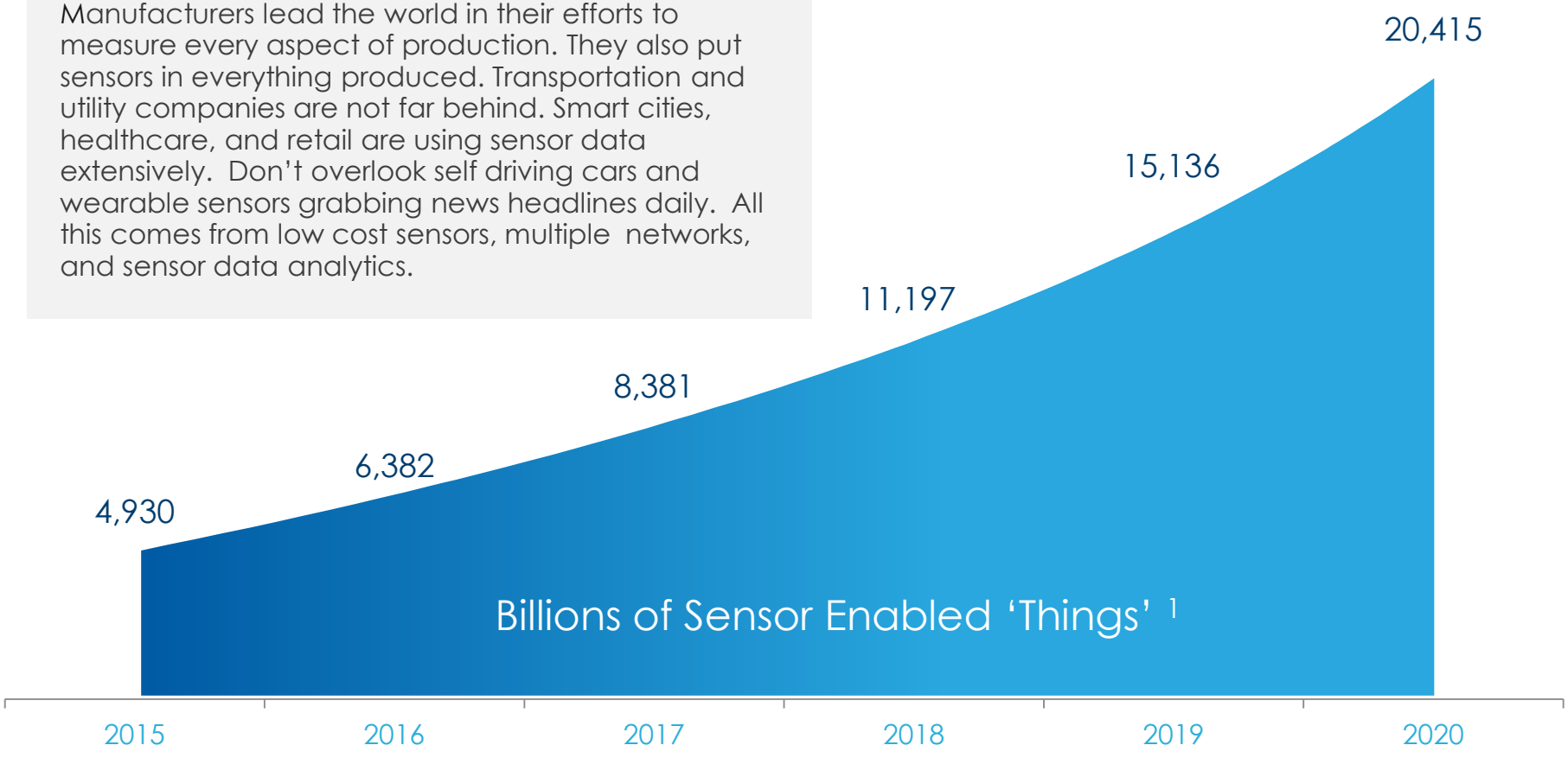
A World Full of Sensors

There and Back Again

Chapter 01

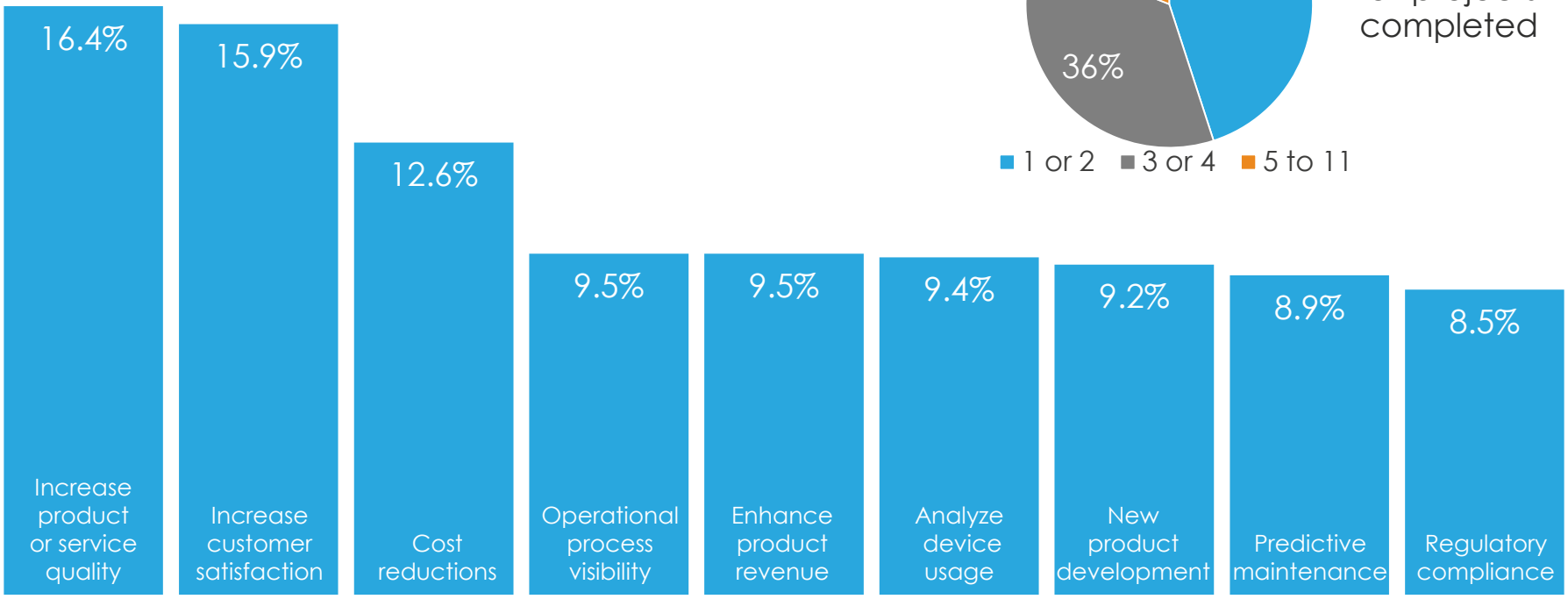
Measuring Everything

Visionaries and fast follower corporations are making giant strides with the Internet of Things (IoT).
Manufacturers lead the world in their efforts to measure every aspect of production. They also put sensors in everything produced. Transportation and utility companies are not far behind. Smart cities, healthcare, and retail are using sensor data extensively. Don't overlook self driving cars and wearable sensors grabbing news headlines daily. All this comes from low cost sensors, multiple networks, and sensor data analytics.



cloudera TERADATA

Business Reasons for IoT



Number of IoT projects completed

Motivations for IoT projects vary by industry, geography, and immediate needs. Of the 250 companies surveyed, 45% are just getting started on their first projects. Thirty-six percent have finished their 3rd or 4th IoT use case. Why are 55% of sites funding three or more IoT projects? Underlying all these business reasons is a solid return on investment. [2]

cloudera TERADATA

Internet of Things (IoT) Concepts

The Edge

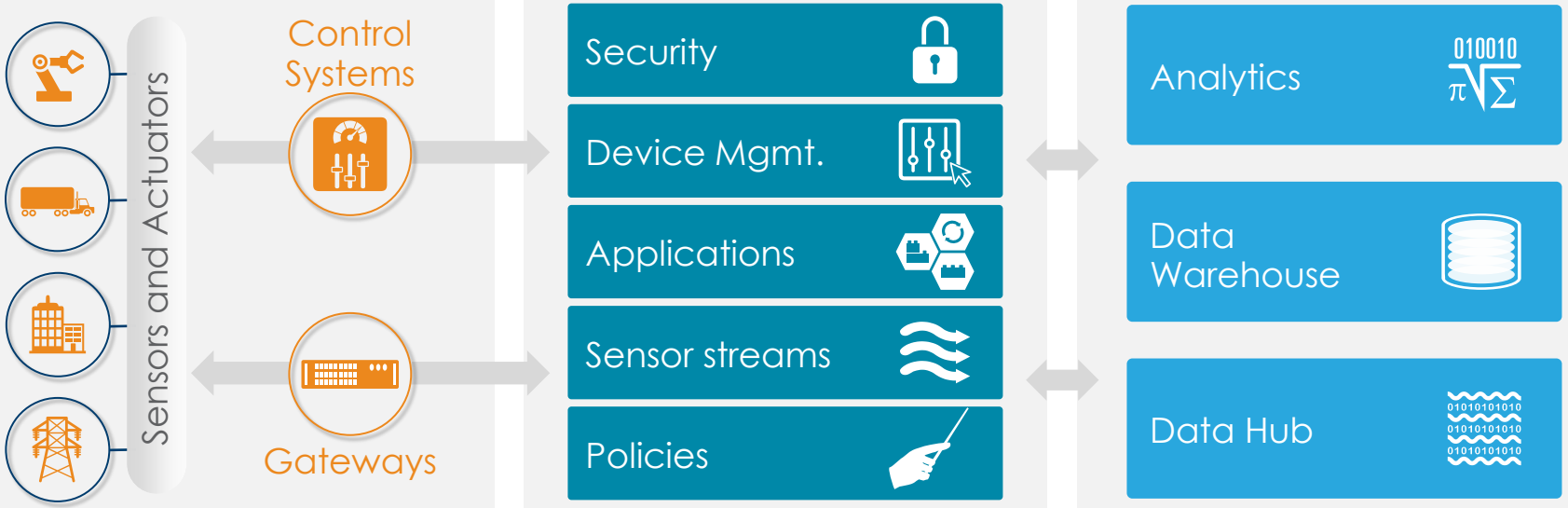
At the edge of the network is Operational Technology where the things are – manufacturing lines, power grids, buildings, and vehicles. Sensors are attached to things, sending messages to the IoT Platform. The edge is part of an intelligent routing data fabric. At the edge, data is acquired, rules applied, models instantiated, routing, and data filtering occurs.

IoT Platform

Here, Information Technology (IT) assists Operational Technology (OT). These daily functions discover new *things*, enroll them in a registry, manage configurations, and respond to alerts. Policies and orchestration automate business rules to manage 1000s of devices. Secure sensor data streams from the edge feed applications and downstream analytics.

Analytics of Things

Data gravity pulls sensor data into huge repositories for deep analysis. The data hub is often the first collection point where raw sensor data is captured and refined. Sensor data is then sent to the data warehouse to add business context and extract more business value. These subsystems generate high return on investment from sensor data.



cloudera TERADATA

Ingestion Streams and Data Preparation Complexity

2 1:12714.609900 2:6.829581 3:2.552121 4:3.394377 5:4.695162 6:-2.268288 7:-3.029891 8:-5.521369 9:15109.250000 10:6.930456 11:3.292046 12:4.469695 13:5.841801 14:-2.664545 15

2 1:13484.328400 2:5.950336 3:3.028627 4:4.075172 5:5.665953 6:-2.487018 7:-3.321967 8:-7.411054 9:16013.197800 10:5.902620 11:3.999665 12:5.602174 13:7.418002 14:-2.951857 15

Unstructured

SID	Date	Time	Value
SS165	6/1/17	1:00:01	6.830
SS165	6/1/17	1:00:02	5.950
SS165	6/1/17	1:00:03	6.494
SS165	6/1/17	1:00:05	6.143
SS170	6/5/17	1:00:01	7.114
SS170	6/5/17	1:00:03	6.194
SS170	6/5/17	1:00:06	7.033

Structured

Sensor data arrives from massively dispersed sources. First, we have to hunt down the things. Automated network searches and discovery of new devices must be part of an IoT Platform. There are too many things out there to rely on manual processes. Sensor data arrives every millisecond, second, minutes, or day. It arrives in batches, bursts, or streams.

While most devices are stationary, many are mobile such as airplanes, cars, and tools. That requires GPS tracking coordinates. It also means data stops arriving the machine goes into tunnels or miles up in the sky. When reconnected to the network, a huge burst of data arrives. A data hub for consistently fast ingest will be mandatory for many companies.

Sensor data formats are also ornery. Every sensor has its own data format –there are no industry standards. It's difficult to integrate IoT data.

Data arrives in XML, JSON, text, or binary. Within the records are date-time stamps and sensor IDs. Plus a huge array of sensor measurements for that sensor. Coping with many sensor formats and metrics is a data transformation hurdle. Then, sensors get upgraded and new data formats appear without notice. Then, of course, sensors sometimes lie. They send bad data that disrupts device operation. Someone first must investigate the data who then sends another someone to fix the sensor. That forces the need for flexibility in parsing and normalizing the data for later use.

Most sensor data is time-series data. If graphed, it looks more like audio waves than rows and columns. Data scientists are often required to restructure the data. They format the streams into arrays, pivot tables, or bins.

cloudera TERADATA

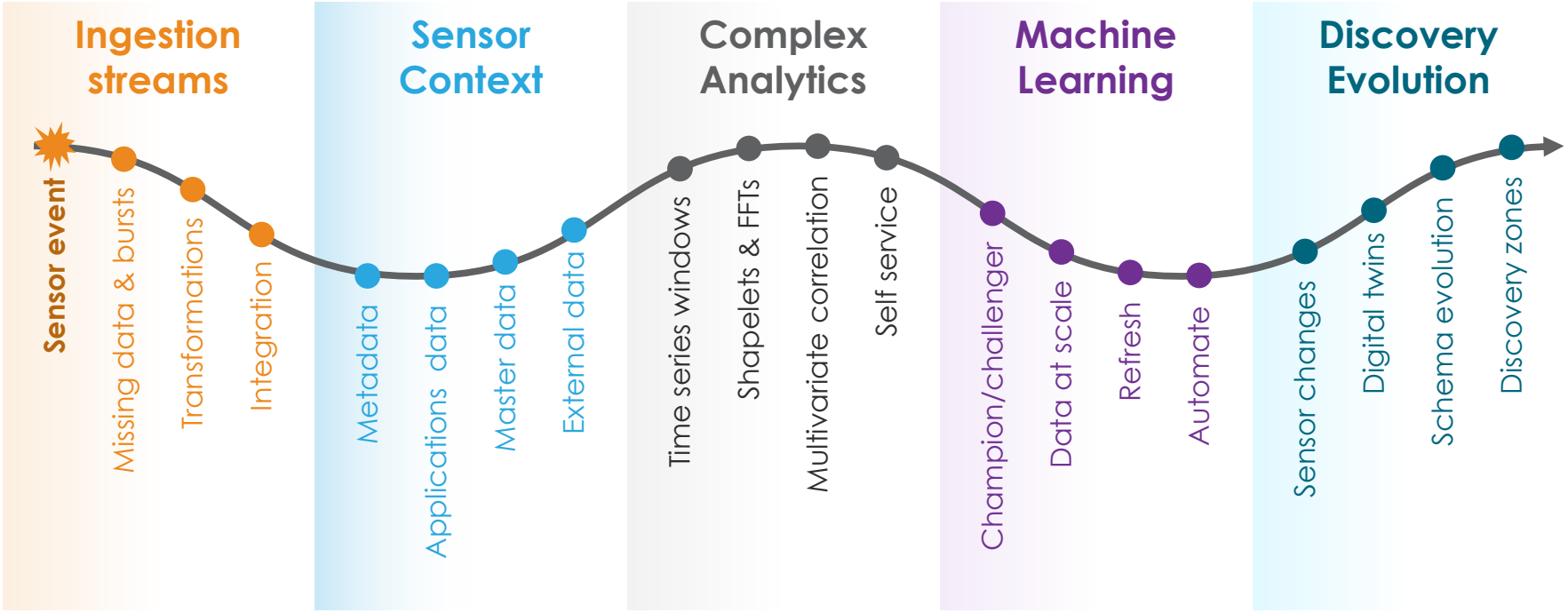
Sensor Data's Journey

There and Back Again

Chapter 02

Sensor Data Journey

From data to insight



Massively dispersed sensor data streams arrive at scale and in bursts.

Sensor data is part of the big picture. Don't leave ROI money on the table.

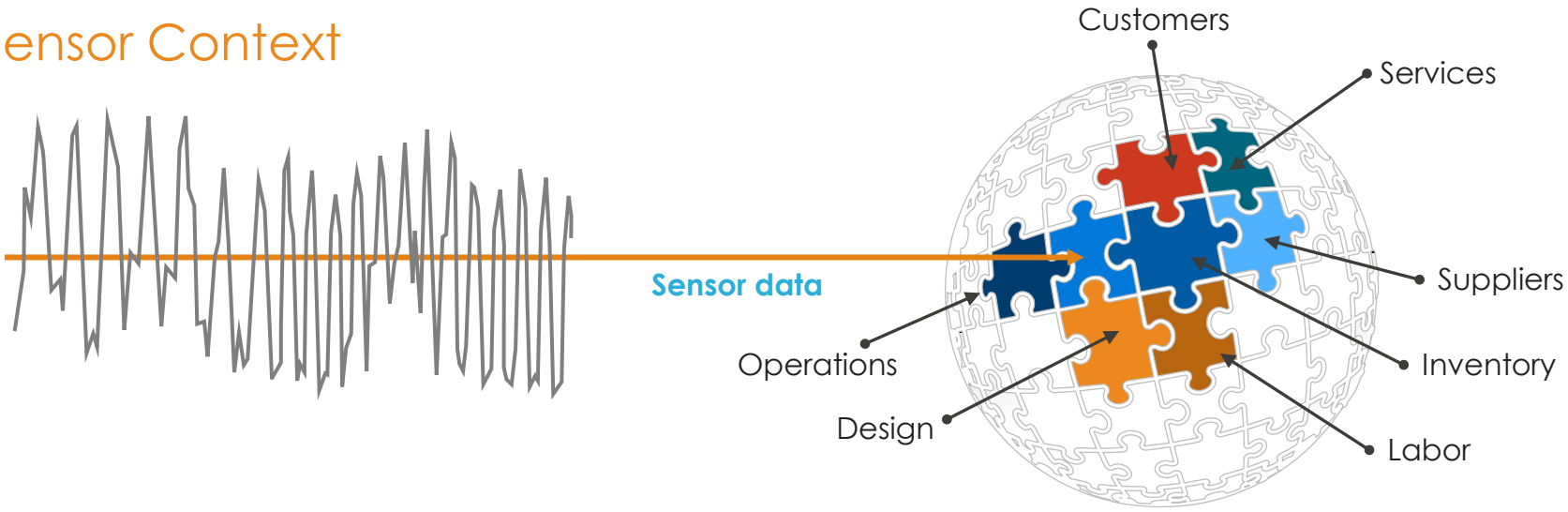
Data scientists use algorithms to create business value. This is not for mortals.

Automating complex problems at scale. And back to the edge.

How will software keep up with device changes?

cloudera TERADATA

Sensor Context



The IoT Platform monitors sensors by their ID, date, time, and geospatial coordinates. This data plus a payload of measurements is in every sensor message sent. The IoT Platform also needs to capture the sensor's version and firmware levels. These facts clarify the structure of data in the message payload. Also important is assessing a sensor's normal data values and arrival frequency. This helps detect when the sensor fails or is erratic. Sensors sometimes lie. Imagine a sensor that shuts down a $500,000 machine by sending bad data. How would we know without all the prior context?

Combinations of sensors provide context too. Like a doctor examining a patient, all the sensors on the 'thing' are considered together. This yields a more complete picture of the health and performance of the device.

There's an even bigger context. Combining sensor data with the business context, its value expands exponentially. Business people and processes make smarter decisions. Imagine a fleet of autonomous trucks and cars. Fleet managers proactively schedule maintenance bays, labor, and parts at a convenient time. Engineers can rethink vehicle designs based on actual usage. They can trace design flaws back to suppliers or unexpected usage. Warranty reserves shrink when the vehicles fail less often and last longer. Insurance premiums drop when fleet usage data shows good driving habits in low risk locations. And customers never face a disruption in service. Thus the sensor data fits into the larger picture of corporate needs and efficiency.

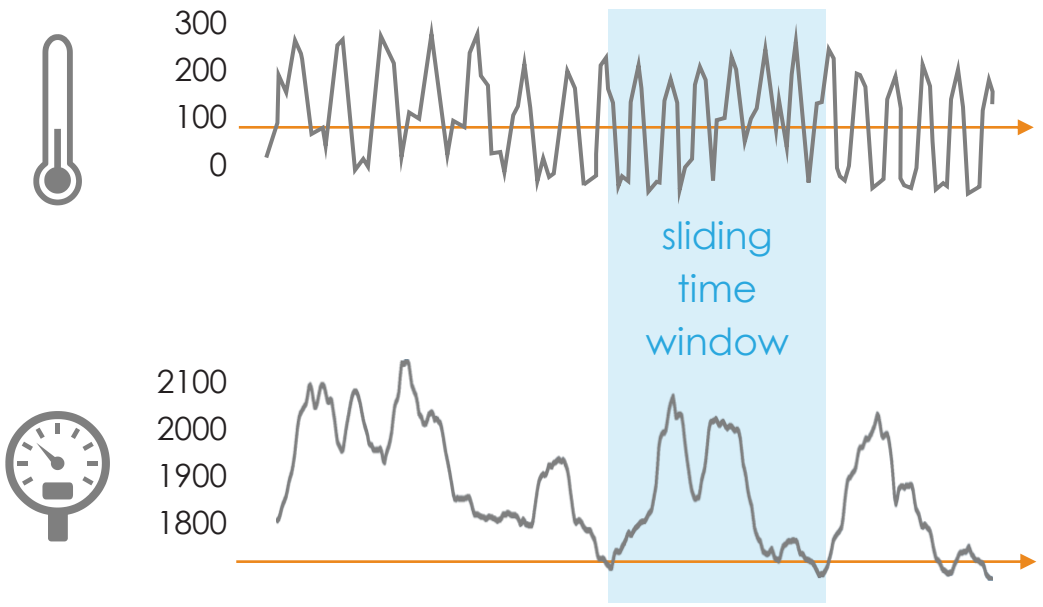
cloudera TERADATA

Complex Analytics

Analytics on sensor data are complex and diverse. Curve fitting algorithms change saw tooth measurements into aggregated values by time intervals. Sliding time windows define the context of when events occur together. Within a time window, multiple advanced algorithms are often applied. Algorithms such as shapelets, Fourier transformations, SAX, and Kahlman filters.

Multivariate correlations help decipher dissimilar sensor values. Imagine dozens of sensors measured in centigrade and dozens more in pressure per square inch. These sensors cannot be directly compared. But we can determine if any of them go above prescribed limits at the same moment in time.

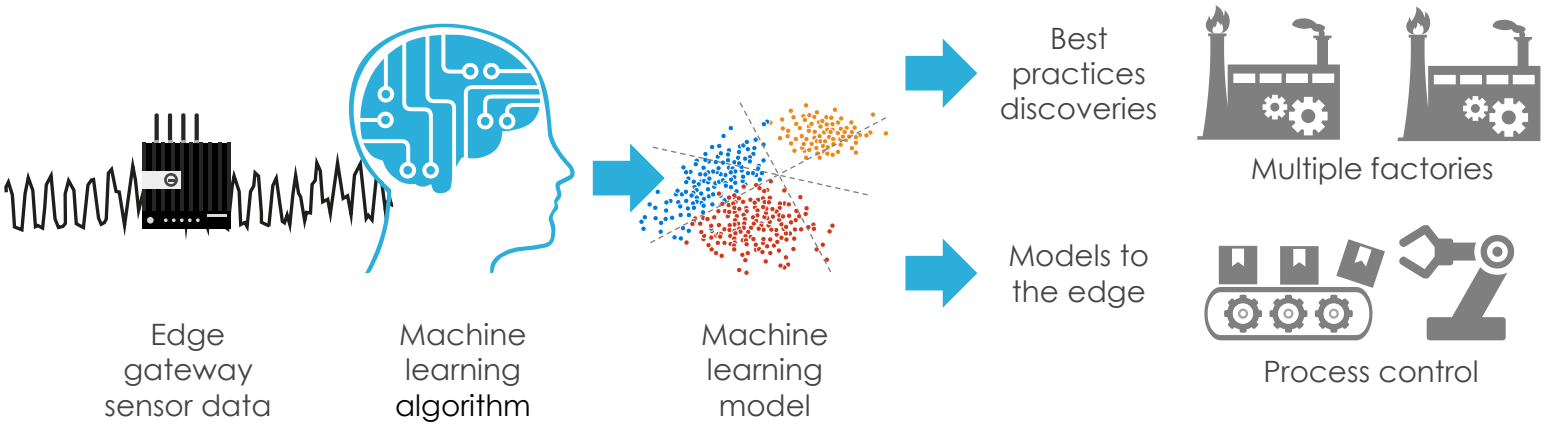
We can't depend exclusively on data scientists to analyze sensor data. Self service BI tools can be used by line of business staff if the analytics are easily invoked. Cloudera's Data Science Workbench is an integrated development environment for programmers. Teradata QueryGrid™ makes joining sensor data in the data hub to the data warehouse easy. Note that data scientists like self-service tools as well.



"Analytics are essential to the success of IoT systems. They are arguably the main point of the IoT as they support the decision-making process in operations that are created in business transformation and digital business programs." [3]

cloudera TERADATA

Machine Learning Models



Machine learning (ML) is the process of feeding algorithms enough data so they can identify a pattern. There's no programming involved. The algorithm outputs the pattern it learned to a small file called a model. The accuracy of a model is dependent on the amount and variety of data fed into the algorithm. *Mo' data is mo' betta*.

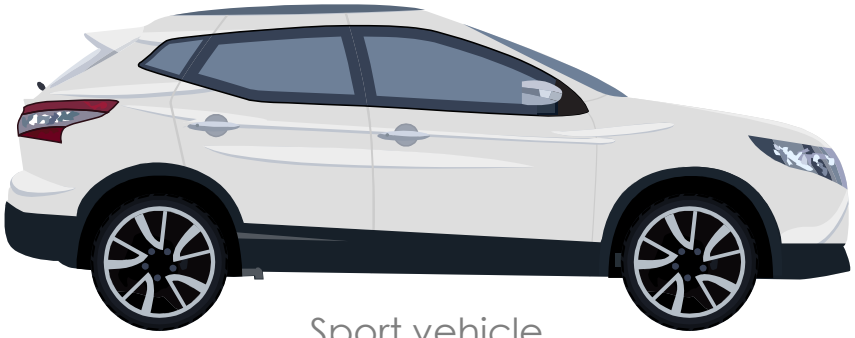
ML models are commonly built using tens to hundreds of terabytes. This is why there's so much demand for hardware and software scalability. For example, one manufacturer keeps more than a petabyte of sensor data in their Hadoop data hub. Machine learning models predict failures, spot outliers, and detect best practices across devices.

Since models are small, they can be deployed in edge gateways. Edge gateways normally use standard if-then-else rules based programming. But there are situations no programmer can solve. Just 10 sensors on a wind turbine sometimes lead to thousands of programming rules. Edge gateways cannot build models but they can use them.

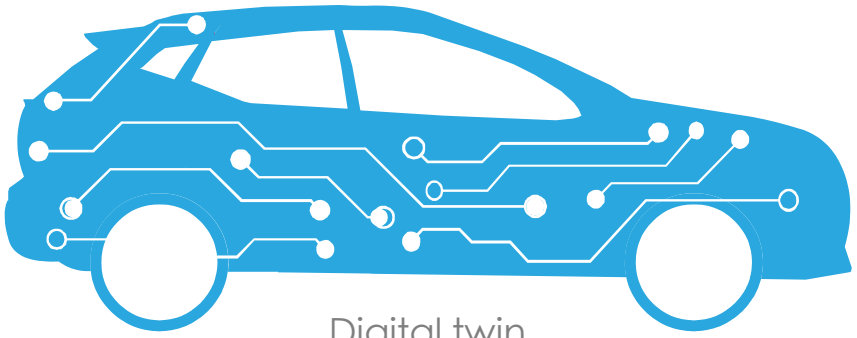
Machine learning models must be recomputed periodically. As new devices and sensors evolve, the model must also evolve. Ideally, this process is automated. A data scientist should periodically do a challenger/champion review of the new model. Only models that perform better than the prior champion replace the prior model production.

cloudera TERADATA

Evolution of Discovery



Sport vehicle



Digital twin

***Things* constantly evolve as new devices and new data formats replace the old.** Failing sensors are replaced with new sensors that emit new data. New *things* are installed with additional sensors. This necessitates schema evolution to keep up.

A digital twin is a model for each *thing* sold or installed. Imagine a digital simulation of each thing installed. That's a digital twin. Digital twins contain a unique identifier and metadata about the device and sensors. They also include time-series data, and contextual data. They can be queried or sent a message to take real world actions. Analytics -- rules or algorithms-- are always used on digital twins. Digital twins enable simulations or modelling of the behavior of real-world things.

Digital twins enable schema evolution and context aggregation. They ease the evolution from, say, condition based maintenance to asset optimization. Thus they engage with the big picture context too.

Start with simple digital twins. Don't get wrapped around the axle with technical fervor. Build a simple schema that can evolve as IoT skills expand.

Data scientist analyzing digital twins need a discovery zone. The data scientist sifts through the data looking for unknown facts. They apply champion/challenger techniques to the digital model to find eureka discoveries. Trial and error tests on the digital twin are later put into practice. Some discoveries will be implemented in the edge computing domain.

The discovery zone spans tools, repositories, and skills. A discovery found in one tool set (i.e. SAS, Aster, or Python) might move to C++ in production. Discoveries are deployed in Hadoop-based data hub, a data warehouse, or a simple script. Moving insights from a discovery zone into production is where many projects fail. Turning data science prototypes into production is crucial skill to nurture.

cloudera TERADATA

Best Fit Engineering

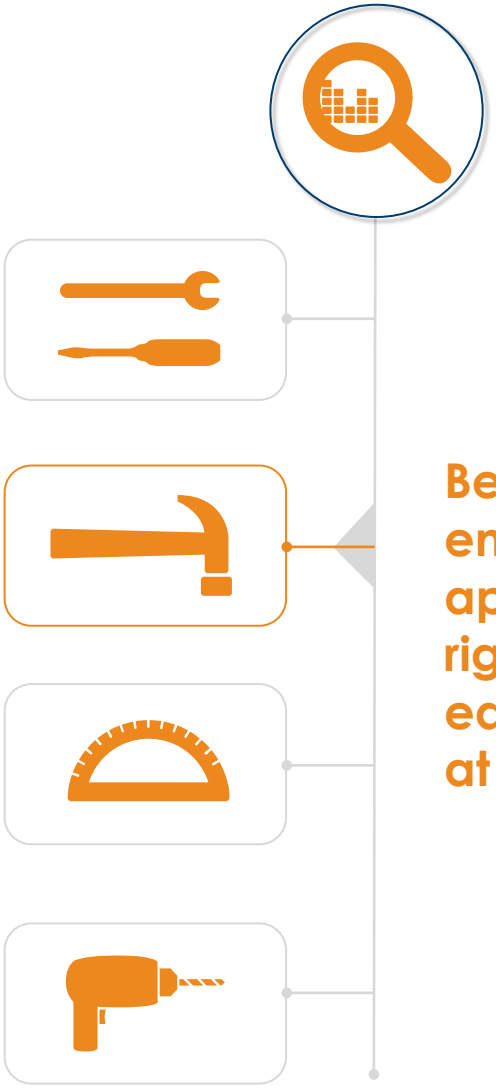
Matching Strengths to Needs

Chapter 03

What is Best Fit Engineering?

Best fit engineering applies the right tool for each problem at hand. Tools and middleware today are different enough that this should be easy. But it's not. It means the software engineers must fully understand the requirements. Then she must map them to the attributes of the software 'parts' available. This approach discards personal bias for favorite tools. Also discarded are single vendor software stack. They simply do not have the best fit components in every category. A well-engineered architecture always has multiple technologies integrated into a single ecosystem. No one tool solves all the problems well.

There are many trade-offs and optimizations to consider. This is especially true when there is some overlap between choices. Factors to consider are the type and size of data, data velocity, and where it's currently located. Data gravity teaches us its best to move processing to the data rather than move data to the processing. This applies to huge data volume but also to IoT edge computing requirements. Also considered is the type of processing and transformations required. And of course who needs to access that data dictates software decisions.



Best fit engineering applies the right tool for each problem at hand.

cloudera TERADATA

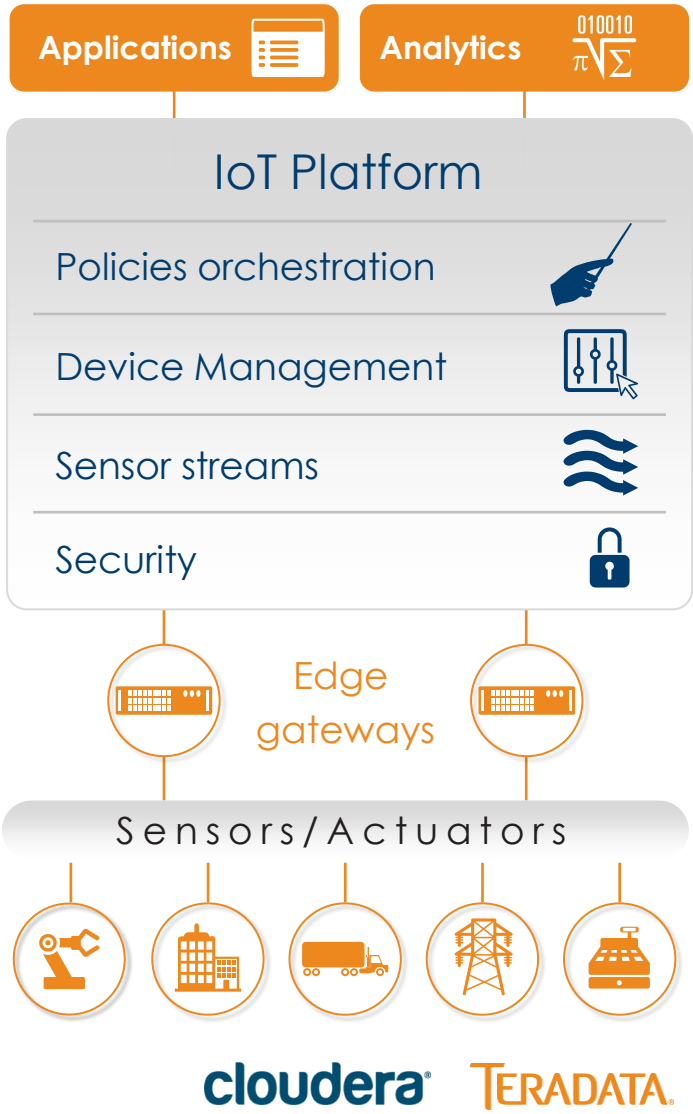
Building the IoT Platform with Open Source

The IoT Platform is the control tower for all the things and sensors. This platform watches the networks, discovering new devices, and monitoring all the sensors. It operates in a modular, distributed multi-tier environment -- i.e. the edge, IoT platform, and enterprise. The IoT platform installs, provisions, and manages IoT endpoints. It securely connects to those IoT endpoints. It ingests sensor data, manages it, and distributes it to analytic subsystems. It enforces policies and rules, orchestrating best use and maintenance of the things. Whew.

Point solutions beguile line of business and operational technology people to build yet-another-data-silo. But when the next IoT projects are planned, point solutions can't adapt. This drives up costs either through rewrites or duplication of infrastructure. Look instead for vendor frameworks with extensive integration features.

Fortunately, open-source components are available for many parts of an IoT solution. That includes edge gateways (e.g. EdgeX Foundry, Eclipse Kura) and IoT Platforms (e.g. OpenIoT, Eclipse Kapua, IoTivity). In many cases, IoT solutions are assembled from open source and existing commercial infrastructure. This means extending and enriching existing systems where your skills are already strong. Half the effort to build an IoT Platform will be integrating components.

For your first IoT Platform effort, engaging an IoT seasoned professional services firm will reduce risks and speed deployment. Hire a sherpa. Start on the right path for the first use cases. And pick something easy for the first IoT project.



Data Hub IoT Strengths

The data hub design pattern begins with scalability at a low cost. The scale of the data can reach petabytes of storage and processing when needed. The cost limitation enables capturing mountains of raw data in its original fidelity and format. The raw data doesn't change, refined derivatives are built. Raw data is retained for long term storage in the data hub, and used again and again. Another part of the data hub pattern is it often supplies data to downstream systems. Thus the name 'hub' indicating a clearing house for all kinds of data. A data hub is often used for high speed ingest, discovery zone work, and refined data distribution.

Some of the key attributes of why Cloudera Enterprise Data Hub lends itself to IoT data management and analytics include:

- Effectively handles multiple IoT data-types, structures, and evolving schemas
- High speed ingest
- Low cost storage
- Real-time processing on IoT streaming data
- Deploy the platform on-premises, in the cloud, or in a hybrid environment based
- Security is at the core
- Provides an agile discovery zone for data scientists or operational technology engineers

cloudera TERADATA

Integrated Data Warehouse IoT Strengths

Sensors | Customers | Inventory | Services

The data warehouse pattern begins with a subject oriented model of the business. Like a digital twin, it reflects the organization's customers, inventory, products, suppliers, and transactions. The subjects in the model are tightly integrated and consistent. This means all data formats and values are standardized. For example, dates, times, account types, and amounts are in consistent formats. This makes it easy for business people to use the data. The data warehouse is persisted, not virtual. Detailed data is retained for historical analysis.

Teradata Database is the cornerstone of many data warehouses. Unique functions useful to IoT design engineers include:

- Special geospatial and time series IoT data services
- Easy self service BI tools
- Hundreds of concurrent users and tasks
- SLAs achieved via workload management
- Fast performance from a cost based optimizer and assorted indexing
- Easy parallelism and scalability to petabytes

Operations | Design | Suppliers | Labor

cloudera TERADATA

Prescriptive Examples

What Visionaries are Doing

Preventing Derailments with Sensors. A Teradata Data Warehouse Example

Derailments are a train companies' worst nightmare. Lives may be lost, shipments destroyed, and recovery costs tens of millions of dollars. Derailments come from overheated wheel bearings lacking enough grease in the axle box. Why?

When the brakes are applied, hundreds of tons of pressure push-back on the track with steel wheels. This causes miniscule flat spots on the wheels the size of a Euro coin. Do it enough times and the wheel becomes slightly out-of-round. This causes axles to heat up. Undetected, the axle box becomes white hot. Then the axle snaps and derails.

This visionary company installed sensors along every 20 miles of track. Trains rush by at 70 miles per hour pulling 20,000 tons of boxcars. Sensors capture infrared temperature readings on the wheel bearings. Other sensors listen to the screeching sound of the wheels. Today, advanced analytics predict wheel bearing failures 8-12 weeks before a derailment.

Dashboards and alerts predict when a wheel is going to fail weeks before it causes a failure. Everyday, the train company spots up to 1500 anomalies out of 20 million sensor readings. Administrators decide within five minutes of a reading to pull a train off the track. Or they may choose to slow it from 70 to 35 mph targeting repairs at the next station. Predictive analytics enables scheduling of parts, labor, and maintenance long before any failures. And that has led to a 75% reduction in derailments.

This means on-time delivery from 3300 trains over 32,000 miles of tracks per day. Today, new sensors are being added along the tracks to gather more information. Next stop: automated driving to reduce the human element in derailments.

cloudera® TERADATA®

Improving Fleet Maintenance Costs. A Cloudera Data Hub Example

Innovations in truck performance and reliability is a strength of this manufacturer. To do this, they analyze over 70 sensor data feeds from 300,000 trucks in real time. Analysis includes predictive analytics, vehicle management, remote diagnostics, and route optimization. Real time analysis monitors acceleration, braking, fuel economy, geolocation, and predicts potential failures. The data is ingested into Cloudera Enterprise for processing and analytics.

Having a truck in the shop costs up to $1,000 per day in lost revenue for vehicle owners. To improve reliability, the manufacturer acquires vehicle usage data from 13 telematics providers. This sensor data is added to the larger business context of geographical metrics, engineering designs, traffic data, warranties, service records, and parts inventories. This allows them to predict issues and send out service advisory's to vehicle owners before problems occur. Service plans and parts are delivered to the nearest dealer when problems do occur.

Fuel costs can account for up to 40% of the total cost of running a fleet. That's why instant feedback is sent to the vehicle driver to improve fuel economy as they are driving. The same sensor data improves their internal testing and customer predictive maintenance.

Need more innovation? Their portal helps customers make informed decisions when ordering trucks. It uses machine learning algorithms to recommend features mapped to requirements.

Owners can track fleet performance from their smartphones or tablets. Better still, vehicle downtime fell by as much as 40 percent.

Smart City Transportation Operations Data Warehouse + Data Hub Example

These city planners support 6.8 million public transportation travelers daily. Their goals are to reduce wait times and traffic. This means enhancing customer journeys by adding routes and optimizing vehicle usage. To do that, they exploit time-series event data with geospatial coordinates.

The city wins awards from Gartner and others for their use of the data warehouse. The data warehouse is focused on analysis of taxis, buses, and train movement. Everyday city planners strive to maximize transportation accessibility and pricing. Popular BI tools are in constant use. Correlation analysis is applied for planning and smooth operations. Increased sensing capabilities track crowd levels in the transit network. Analytics also forecast congestion caused by special events or train incidents. And no personal identifiable data is found on their open data crowdsourcing website.

In 2016, they acquired a Teradata Aster®/Cloudera appliance. Wipro was chosen as their system integrator. With Wipro's help, sensor data from buses began flowing into HDFS using Kafka and Flume. Spark jobs were setup to aggregate the data which goes into Hive for user queries. Now, for each bus stop, traffic planners look deeper into whether a bus is on time and if there are too many people on the bus. They know precisely when they need more capacity.

New sensor data streams arrive from 28,000 taxis in five operator fleets. Streams of GIS data flow into the system all day long. GIS analysis reveals regions of clustered taxi usage by street and time of day. Distance traveled per trip is also analyzed in the context of traffic congestion.

Wipro received the "Best Collaboration Partner" award from the city in 2016 for their work on the transportation data warehouse.

cloudera TERADATA

Conclusions

The Analytics of Things is inevitable. The first use of sensor data is operational – simply monitoring and adjusting the operation of devices. But tactical and strategic analysis is where the ROI payback shines brightest. Visionary corporations have already done 5-10 IoT projects. Yet there's still plenty of time to be a fast follower.

Sensor data adds new ingest and analysis requirements. But sensors add high velocity ingest from hundreds of sources. Deploy advanced analytics to convert continuous data streams into decisions. Machine learning applies the higher order analysis to the things analyzed. Start by exploiting existing big data analytic skills and tools.

Strong data integration and data quality strengthen IoT deployments and results. Data governance and security should always be top of mind as well.

Apply best-fit engineering to tasks and data flows. IoT architects must optimize trade-offs between components. Many products -- open source, commercial, and do-it-yourself -- feed into this optimization.

Hire IoT savvy consultants for your first IoT project. Learn quickly but don't get trapped in never ending proof of concept mode. Use agile development methods with line of business and operational technology staff on the team.

About Cloudera

Cloudera delivers the modern platform for machine learning and advanced analytics built on the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems by efficiently capturing, storing, processing and analyzing vast amounts of data. Learn more at cloudera.com.

About Teradata

Teradata empowers companies to achieve high-impact business outcomes through analytics. With a powerful combination of Industry expertise and leading hybrid cloud technologies for data warehousing and big data analytics, Teradata unleashes the potential of great companies.

EB9830 09.17

Endnotes:

1. Gartner, Forecast: Internet of Things — Endpoints and Associated Services, Worldwide, December 2016
2. EMA, The Internet of Things – How to Get There from Here, March 2017
3. Gartner, Three Best Practices for Internet of Things Analytics, October 23, 2015