SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY

UNIT – II Notes

Big Data Technologies: Hadoop's Parallel World – Data discovery – Open source technology for Big Data Analytics – cloud and Big Data –Predictive Analytics – Mobile Business Intelligence and Big Data

Big Data Technologies: Hadoop's Parallel World

Brief History of Hadoop

There are many Big Data technologies that have been making an impact on the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk.

→Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.

→The original creators of Hadoop are Doug Cutting (used to be at Yahoo! now at Cloudera) and Mike

→Doug and Mike were building a project called "Nutch" with the goal of creating a large Web index.

→They saw the MapReduce and GFS papers from Google, which were obviously super relevant to the problem Nutch was trying to solve.

→Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve:

The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.

→The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before, at a transformationally lower TCO than proprietary solutions and thereby take advantage of all their data to increase operational efficiency and gain a competitive edge.

→At one-tenth the cost of traditional solutions, Hadoop excels at supporting complex analyses— including detailed, special-purpose computation—across large collections of data.

Hadoop workloads

Hadoop handles a variety of workloads, including search, log processing, recommendation systems, data warehousing, and video/image analysis.

Today 's explosion of data types and volumes means that Big Data equals big opportunities and Apache Hadoop empowers organizations to work on the most modern scale-out architectures using a clean-sheet design data framework, without vendor lock-in.

- Apache Hadoop is an open-source project administered by the Apache Software Foundation.
- The software was originally developed by the world 's largest Internet companies to capture and analyze the data that they generate.
- Unlike traditional, structured platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data.
- Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system fail.
- Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

The two critical components of Hadoop are:

1. The Hadoop Distributed File System (HDFS).

HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

2. MapReduce.

Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

Features of Hadoop

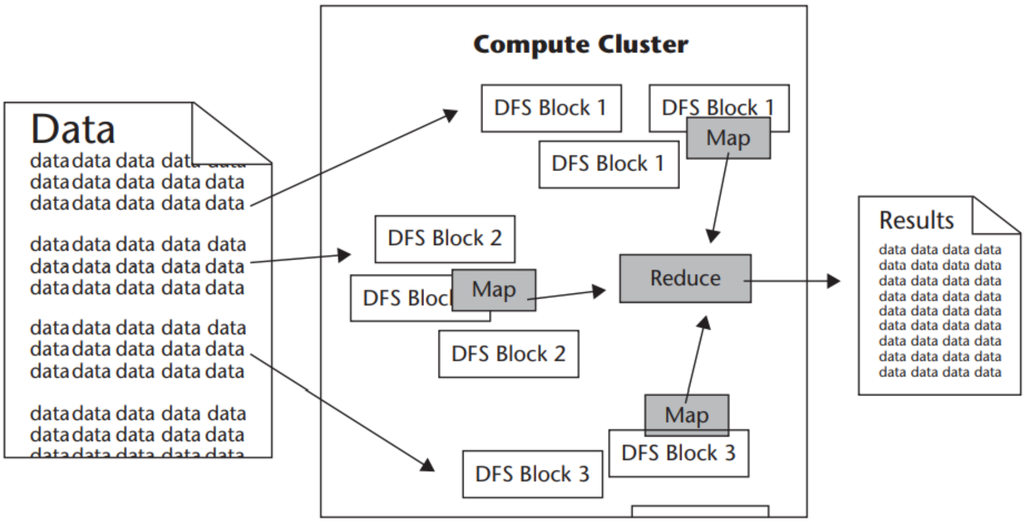
→Both HDFS and MapReduce are designed to continue to work in the face of system failures.

→HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster.

→Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job.

→If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data.

→Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost.



Source: Apache Software Foundation.

Old vs. New Approaches

- The old way is a data and analytics technology stack with different layers “cross-communicating data” and working on “scale-up” expensive hardware.
- The new way is a data and analytics platform that does all the data processing and analytics in one “layer,” without moving data back and forth on cheap but scalable (“scale out”) commodity hardware.

- The new approach is based on two foundational concepts. Number one, data needs to be stored in a system in which the hardware is infinitely scalable.

- In other words, you cannot allow hardware (storage and network) to become the bottleneck. Number two, data must be processed, and converted into usable business intelligence where it sits.

  The three main important points are

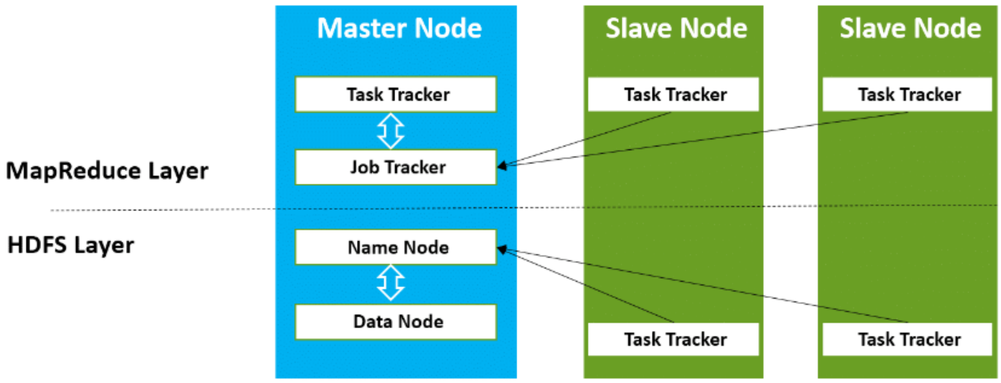
  1.The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data.

  2. Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing.

  3. The variety of data is on the rise and the ability to handle unstructured data is on the rise.

The Hadoop High-level Architecture

Hadoop Architecture based on the two main components namely MapReduce and HDFS



The Apache Hadoop Module

Hadoop Common: Includes the common utilities which supports the other Hadoop modules

HDFS: Hadoop Distributed File System provides unrestricted, high-speed access to the data application.

Hadoop YARN: This technology is basically used for scheduling of job and efficient management of the cluster resource.

MapReduce: This is a highly efficient methodology for parallel processing of huge volumes of data.

Apache Ambari : It is a tool for managing, monitoring and provisioning of the Hadoop clusters. Apache Ambari supports the HDFS and MapReduce programs. Major highlights of Ambari are:

- Managing of the Hadoop framework is highly efficient, secure and consistent.
- Management of cluster operations with an intuitive web UI and a robust API
- The installation and configuration of Hadoop cluster are simplified effectively.
- It is used to support automation, smart configuration and recommendations
- Advanced cluster security set-up comes additional with this tool kit.
- The entire cluster can be controlled using the metrics, heat maps, analysis and troubleshooting
- Increased levels of customization and extension make this more valuable.

Cassandra: It is a distributed system to handle extremely huge amount of data which is stored across several commodity servers. The database management system (DBMS)is highly available with no single point of failure.

HBase: it is a non-relational, distributed database management system that works efficiently on sparse data sets and it is highly scalable.

Apache Spark: This is highly agile, scalable and secure the Big Data compute engine, versatiles the sufficient work on a wide variety of applications like real-time processing, machine learning, ETL and so on.

Hive: It is a data warehouse tool basically used for analyzing, querying and summarizing of analyzed data concepts on top of the Hadoop framework.

Pig: Pig is a high-level framework which ensures us to work in coordination either with Apache Spark or MapReduce to analyze the data. The language used to code for the frameworks are known as Pig Latin.

Sqoop: This framework is used for transferring the data to Hadoop from relational databases. This application is based on a command-line interface.

Oozie: This is a scheduling system for workflow management, executing workflow routes for successful completion of the task in a Hadoop.

Zookeeper: Open source centralized service which is used to provide coordination between distributed applications of Hadoop. It offers the registry and synchronization service on a high level.

- **Hadoop Mapreduce (Processing/Computation layer) –**MapReduce is a parallel programming model mainly used for writing large amount of data distribution applications devised from Google for efficient processing of large amounts of datasets, on large group of clusters.
- **Hadoop HDFS (Storage layer) –**Hadoop Distributed File System or HDFS is based on the Google File System (GFS) which provides a distributed file system that is especially designed to run on commodity hardware. It reduces the faults or errors and helps incorporate low-cost hardware. It gives high level processing throughput access to application data and is suitable for applications with large datasets.
- **Hadoop YARN –**Hadoop YARN is a framework used for job scheduling and cluster resource management.
- **Hadoop Common –**This includes Java libraries and utilities which provide those java files which are essential to start Hadoop.
- **Task Tracker –**It is a node which is used to accept the tasks such as shuffle and Mapreduce form job tracker.
- **Job Tracker –**It is a service provider which runs Mapreduce jobs on cluster.
- **Name Node –**It is a node where Hadoop stores all file location information(data stored location) in Hadoop distributed file system.

- **Data Node –** The data is stored in the Hadoop distributed file system.

Data Discovery: Work the Way People 's Minds Work

There is a lot of buzz in the industry about data discovery.

The term used to describe the new wave of business intelligence that enables users to explore data, make discoveries, and uncover insights in a dynamic and intuitive way versus predefined queries and preconfigured drill-down dashboards.

Tableau Software and QlikTech are the two Business intelligence tools used for reporting .

- ➢ Analytics and reporting are produced by the people using the results. IT provides the infrastructure, but business people create their own reports and dashboards.
- ➢ There is a simple example of powerful visualization that the Tableau team is referring to.
- ➢ A company uses an interactive dashboard to track the critical metrics driving their business.

Example of Tableau Software

- A company uses an interactive dashboard to track the critical metrics driving their business.
- Every day, the CEO and other executives are plugged in real-time to see how their markets are performing in terms of sales and profit, what the service quality scores look like against advertising investments, and how products are performing in terms of revenue and profit.
- Interactivity is key: a click on any filter lets the executive look into specific markets or products.
- She can click on any data point in any one view to show the related data in the other views.
- Hovering over a data point can get any unusual pattern or outlier by showing details on demand.

Or she can click through the underlying information in a split-second

"Business intelligence needs to work the way people 's minds work.

Users need to navigate and interact with data any way they want to—asking and answering questions on their own and in big groups or teams.

One capability that we have all become accustomed to is search, what many people refer to as "Googling."

This is a prime example of the way people 's minds work. Qliktech has designed a way for users to leverage direct— and indirect—search.

- With QlikView search, users type relevant words or phrases in any order and get instant, associative results.
- With a global search bar, users can search across the entire data set. With search boxes on individual list boxes, users can confine the search to just that field.
- Users can conduct both direct and indirect searches. For example, if a user wanted to identify a sales rep but couldn't 't remember the sales rep 's name—just details about the person, such as that he sells fish to customers in the Nordic region.

Open-Source Technology for Big Data Analytics

- Open-source big data analytics makes use of open-source software and tools in order to execute big data analytics by either using an entire software platform or various open-source tools for different tasks in the process of data analytics.
- Apache Hadoop is the most well-known system for big data analytics, but other components are required before a real analytics system can be put together.
- Hadoop is the open-source implementation of the MapReduce algorithm pioneered by Google and Yahoo, so it is the basis of most analytics systems today.
- Many big data analytics tools make use of open source, including robust database systems such as the open-source MongoDB, a sophisticated and scalable NoSQL database very suited for big data applications, as well as others.

Open-source big data analytics refers to the use of open-source software and tools for analysing huge quantities of data in order to gather relevant and actionable information that an organization can use in order to further its business goals. The biggest player in open-source big data analytics is Apache's Hadoop – it is the most widely used software library for processing enormous data sets across a cluster of computers using a distributed process for parallelism.

Open-source big data analytics services encompass:

- Data collection system
- Control centre for administering and monitoring clusters
- Machine learning and data mining library
- Application coordination service
- Compute engine
- Execution framework

Proprietary Software:

- Proprietary software is computer software where the source codes are publicly not available only the company that has created them can modify it.
- Here the software is developed and tested by the individual or organization by which it is owned not by the public.
- This software is managed by a closed team of individuals or groups that developed it.
- We have to pay to get this software and its commercial support is available for maintenance.
- The company gives a valid and authenticated license to the users to use this software. But this license puts some restrictions on users also like.
- **Some examples of Proprietary software include Windows, macOS, Internet Explorer, Google Earth, Microsoft Office, etc.**

1. Hadoop

Even if you are a beginner in this field, we are sure that this is not the first time you've read about Hadoop. It is recognized as one of the most popular big data tools to analyze large data sets, as the platform can send data to different servers. Another benefit of using Hadoop is that it can also run on a cloudinfrastructure.

This open-source software framework is used when the data volume exceeds the available memory. This big data tool is also ideal for data exploration, filtration, sampling, and summarization. It consists of four parts:

- **Hadoop Distributed File System:** This file system, commonly known as HDFS, is a distributed file system compatible with very high-scale bandwidth.
- **MapReduce:** It refers to a programming model for processing big data.

- **YARN:** All Hadoop's resources in its infrastructure are managed and scheduled using this platform.
- **Libraries:** They allow other modules to work efficiently with Hadoop.

2. Apache Spark

The next hype in the industry among big data tools is Apache Spark. the reason behind this is that this open-source big data tool fills the gaps of Hadoop when it comes to data processing. This big data tool is the most preferred tool for data analysis over other types of programs due to its ability to store large computations in memory**.** It can run complicated algorithms, which is a prerequisite for dealing with large data sets.

Proficient in handling batch and real-time data, Apache Spark is flexible to work with HDFS and OpenStack Swift or Apache Cassandra. Often used as an alternative to MapReduce, Spark can run tasks 100x faster than Hadoop's MapReduce.

3. Cassandra

Apache Cassandra is one of the best big data tools to process structured data sets. Created in **2008 by Apache Software Foundation,** it is recognized as the best open-source big data tool for scalability. This big data tool has a proven fault-tolerance on cloud infrastructure and commodity hardware, making it more critical for big data uses.

It also offers features that no other relational and NoSQL databases can provide. This includes simple operations, cloud availability points, performance, and continuous availability as a data source, to name a few. Apache Cassandra is used by giants like **Twitter, Cisco, and Netflix.**

4. MongoDB

MongoDB is an ideal alternative to modern databases. A document-oriented database is an ideal choice for businesses that need fast and real-time data for instant decisions. One thing that sets it apart from other traditional databases is that it makes use of documents and collections instead of rows and columns.

Thanks to its power to store data in documents, it is very flexible and can be easily adapted by companies. It can store any data type, be it integer, strings, Booleans, arrays, or objects. MongoDB is easy to learn and provides support for multiple technologies and platforms.

5.Apache Hive

Hive is an open source big data software tool. It allows programmers analyze large data sets on Hadoop. It helps with querying and managing large datasets real fast.

Features:

- It Supports SQL like query language for interaction and Data modeling
- It compiles language with two main tasks map, and reducer
- It allows defining these tasks using Java or Python
- Hive designed for managing and querying only structured data
- Hive's SQL-inspired language separates the user from the complexity of Map Reduce programming
- It offers Java Database Connectivity (JDBC) interface.

6.kaggle

- Kaggle is the world's largest big data community. It helps organizations and researchers to post their data & statistics. It is the best place to analyze data seamlessly.

Features:

- The best place to discover and seamlessly analyze open data
- Search box to find open datasets
- Contribute to the open data movement and connect with other data enthusiasts

7.Apache Hbase

Apache HBase is an open-source, NoSQL, distributed big data store. It enables random, strictly consistent, real-time access to petabytes of data. HBase is very effective for handling large, sparse datasets.

HBase integrates seamlessly with Apache Hadoop and the Hadoop ecosystem and runs on top of the Hadoop Distributed File System (HDFS) or Amazon S3 using Amazon Elastic MapReduce (EMR) file system, or EMRFS. HBase serves as a direct input and output to the Apache MapReduce framework for Hadoop, and works with Apache Phoenix to enable SQL-like queries over HBase tables.

8.Apache Storm

It is a free big data open-source computation system. It is one of the best big data tools that offers a distributed, real-time, fault-tolerant processing system.

Having been benchmarked as processing one million 100-byte messages per second per node, it has big data technologies and tools that use parallel calculations that can run across a cluster of machines.

Being open source, robust and flexible, it is preferred by medium and large-scale organizations. It guarantees data processing even if the messages are lost, or nodes of the cluster die.

9.Apache Pig

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using PIG can also be achieved using java used in MapReduce.

10.Apache Flink

Apache Flink is an open-source platform that provides a scalable, distributed, fault-tolerant, and stateful stream processing capabilities. Flink is one of the most recent and pioneering Big Data processing frameworks.

Apache Flink allows to ingest massive streaming data (up to several terabytes) from different sources and process it in a distributed fashion way across multiple nodes, before pushing the derived streams to other services or applications such as Apache Kafka, DBs, and Elastic search. Simply, the basics building blocks of a Flink pipeline: input, processing, and output. Its runtime supports low-latency processing at extremely high throughputs in a fault-tolerant manner. Flink capabilities enable real-time insights from streaming data and event-based capabilities. Flink enables real-time data analytics on streaming data and fits well for continuous Extract-transform-load (ETL) pipelines on streaming data and for event-driven applications as well.

The open-source projects are managed and supported by commercial companies, such as Cloudera, that provide extra capabilities, training, and professional services that support open-source projects such as Hadoop.

This is similar to what Red Hat has done for the open-source project Linux.

The advantage of the open source stack—flexibility, extensibility, and lower cost.

“One of the great benefits of open source lies in the flexibility of the adoption model: you download and deploy it when you need it,” .With open source, you can try it and adopt it at your own pace.

The several assumptions of big data are

1.The amounts of data generated would be manageable

2. Programming resources would remain scarce

3. Faster data processing would require bigger, more expensive hardware.

- The old model was top-down, slow, inflexible and expensive.
- The new software development model is bottom-up, fast, flexible, and considerably less costly.

- A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors.
- It reflects the old command-and-control mentality of the traditional corporate world and the old economic order.

David then makes the case for an open-source analytics stack.

An open-source stack is defined by its community of users and contributors.

No one "controls" an open-source stack, and no one can predict exactly how it will evolve.

The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

It have been designing their solutions to plug and play with technology such as Hadoop.

For example, Teradata Aster designed SQL-H, which is a seamless way to execute SQL and SQL-MapReduce on Apache Hadoop data.

Cloud and Big Data

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). It's a virtualization framework.

It is like a resource on demand whether it be storage, computing etc. Cloud follows pay per usage model. You need to pay the amount of resource you use.

This computing service by cloud charges you based only on the amount of computing resources we use. So for example, if you want to give demo to a client on a cluster of more than 100 machines and you do not have so many machines currently available with you, then in such case cloud computing plays a very important role.

Cloud plays an important role within the Big Data world, by providing horizontally expandable and optimized infrastructure that supports practical implementation of Big Data.

In cloud computing, all data is gathered in data centers and then distributed to the end-users. Further, automatic backups and recovery of data is also ensured for business continuity, all such resources are available in the cloud.

We do not know exact physical location of these resources provided to us. You just need dummy terminals like desktops, laptops, phones etc. and a net connection.

There are multiple ways to access the cloud:

1. Applications or software as a service (SAAS) ex. Salesforce.com, dropbox, google drive etc.
2. Platform as a service (PAAS)
3. Infrastructure as a service (IAAS)

Features of Cloud Computing

Let us see few features of cloud computing:

a. Scalability

Scalability is provided by using distributed computing

b. Elasticity

Customers are allowed to use and pay for only that much resource which it is using.

In cloud computing, elasticity is defined as the degree to which a system is able to adapt to workload changes in an autonomic manner, so that at any time the available resources match the current demand as closely as possible.

c. Resource Pooling

Same resources are allowed to be used by multiple organizations. The computing resources are pooled for serving various consumers via multi-tenant model, with different resources dynamically assigned and reassigned according to consumer demand.

d. Self service

Customers are provided easy to use interface through which they can choose services they want. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed without requiring human interaction.

e. Low Costs

It charges you based only on the amount of computing resources we use and you need not buy expensive infrastructure. Pricing on a utility computing basis is usage-based and fewer IT skills are required for implementation.

f. Fault Tolerance

Allows recovery in case of a part in cloud system fails to respond.

Cloud Deployment Models

There are mainly 2 types of cloud deployments models:

- Public cloud – A cloud is called a “public cloud” when the services are open over a network for public use.
- Private Cloud – Private cloud is operated solely for a single organization, whether managed internally or by a third-party, and hosted either internally or externally.

Cloud Delivery Models

Cloud services are categorized as below:

1. **Infrastructure as a service (IAAS):** It means complete infrastructure will be provided to you. Maintenance related tasks will be done by cloud provider and you can use it as per your requirement. It can be used as public and private both.

Examples of IaaS are virtual machines, load balancers, and network attached storage.

2. **Platform as a service (PAAS):** Here we have object storage, queuing, databases, runtime etc. All these we can get directly from the cloud provider. It's our responsibility to configure and use that. Providers will give us the resources but connectivity to our database and other similar activities are our responsibility.

Examples of PaaS are Windows Azure and Google App Engine (GAE).

3. **Applications or software as a service** (SAAS) ex. Salesforce.com, dropbox, google drive etc. Here we do not have any responsibility. We are using the application that is running on the cloud. All infrastructure setup is the responsibility of the service provider.

For SaaS to work, the infrastructure (IaaS) and the platform (PaaS) must be in place.

Cloud for Big Data

Below are some examples of how cloud applications are used for Big Data:

IAAS in a public cloud: Using a cloud provider's infrastructure for Big Data services, gives access to almost limitless storage and compute power.
IaaS can be utilized by enterprise customers to create cost-effective and easily scalable IT solutions where cloud providers bear the complexities and expenses of managing the underlying hardware.

If the scale of a business customer's operations fluctuates, or they are looking to expand, they can tap into the cloud resource as and when they need it rather than purchase, install and integrate hardware themselves.

PAAS in a private cloud: PaaS vendors are beginning to incorporate Big Data technologies such as Hadoop and MapReduce into their PaaS offerings, which eliminate the dealing with the complexities of managing individual software and hardware elements.
For example, web developers can use individual PaaS environments at every stage of development, testing and ultimately hosting their websites. However, businesses that are developing their own internal software can also utilize Platform as a Service, particularly to create distinct ring-fenced development and testing environments.

SAAS in a hybrid cloud: Many organizations feel the need to analyze the customer's voice, especially on social media. SaaS vendors provide the platform for the analysis as well as the social media data.

Office software is the best example of businesses utilizing SaaS. Tasks related to accounting, sales, invoicing, and planning can all be performed through SAAS. Businesses may wish to use one piece of software that performs all of these tasks or several that each performs different tasks.

The software can be subscribed through the internet and then accessed online via any computer in the office using a username and password. If needed, they can switch to software that fulfills their requirements in better manner.

Everyone who needs access to a particular piece of software can be set up as a user, whether it is one or two people or every employee in a corporation that employs hundreds.

Providers in the Big Data Cloud Market

In addition there are many startups that have interesting products in cloud space. Here we have a list of major vendors of cloud computing. Few of the cloud providers are google, citrix, netmagic, redhat, rackspace etc. Amazon (aws) is the leading cloud provider amongst all. Microsoft is also providing cloud services and it is called as azure.

Infrastructure as a Service cloud computing companies:

- Amazon's offerings include S3 (Data storage/file system), SimpleDB (non-relational database) and EC2 (computing servers).
- Rackspace's offerings include Cloud Drive (Data storage/file system), Cloud Sites (web site hosting on cloud) and Cloud Servers(computing servers).
- IBM's offerings include Smart Business Storage Cloud and Computing on Demand (CoD).
- AT&T's provides Synaptic Storage and Synaptic Compute as a service.

Platform as a Service cloud computing companies

- Googles AppEngine is a development platform that is built upon Python and Java.
- com's provides a development platform that is based upon Apex.
- Microsoft Azure provides a development platform based upon .Net.

Software as a Service companies

- In SaaS, Google provides space that includes Google Docs, Gmail, Google Calendar and Picasa.
- IBM provides LotusLive iNotes, a web-based email service for messaging and calendaring capabilities to business users.

- Zoho provides online products similar to Microsoft office suite.

Predictive Analytics Moves into the Limelight

What is predictive analyticts?

Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning. Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. Predictive analytics is often associated with big data and data science.

predictive analytics models are designed to assess historical data, discover patterns, observe trends, and use that information to predict future trends.

Predictive analytics can be deployed in across various industries for different business problems. Below are a few industry use cases to illustrate how predictive analytics can inform decision-making within real-world situations.

- **Banking:** Financial services use machine learning and quantitative tools to predict credit risk and detect fraud. Predictive analytics allows them to support dynamic market changes in real-time in addition to static market constraints. This use of technology allows it to both customize personal services for clients and to minimize risk.
- **Healthcare:** Predictive analytics in health care is used to detect and manage the care of chronically ill patients, as well as to track specific infections such as sepsis. Geisinger Health used predictive analytics to mine health records to learn more about how sepsis is diagnosed and treated. Geisinger created a predictive model based on health records for more than 10,000 patients who had been diagnosed with sepsis in the past. The model yielded impressive results, correctly predicting patients with a high rate of survival.
- **Human resources (HR):** HR teams use predictive analytics and employee survey metrics to match prospective job applicants, reduce employee turnover and increase employee engagement. This combination of quantitative and qualitative data allows businesses to reduce their recruiting costs and increase employee satisfaction, which is particularly useful when labor markets are volatile.
- **Marketing and sales:** While marketing and sales teams are very familiar with business intelligence reports to understand historical sales performance, predictive analytics enables companies to be more proactive in the way that they engage with their clients across the customer lifecycle. For example, churn predictions can enable sales teams to identify dissatisfied clients sooner, enabling them to initiate conversations to promote retention. Marketing teams can leverage predictive data analysis for cross-sell strategies, and this commonly manifests itself through a recommendation engine on a brand's website.
- **Supply chain:** Businesses commonly use predictive analytics to manage product inventory and set pricing strategies. This type of predictive analysis helps companies meet customer demand without overstocking warehouses. It also enables companies to assess the cost and return on their products over time. If one part of a given product becomes more expensive to import, companies can project the long-term impact on revenue if they do or do not pass on additional costs to their customer base. For a deeper look at a case study, you can read more about how FleetPride used this type of data

analytics to inform their decision making on their inventory of parts for excavators and tractor trailers. Past shipping orders enabled them to plan more precisely to set appropriate supply thresholds based on demand.

Benefits of predictive modelling

- **Security:** Every modern organization must be concerned with keeping data secure. A combination of automation and predictive analytics improves security. Specific patterns associated with suspicious and unusual end user behavior can trigger specific security procedures.
- **Risk reduction:** In addition to keeping data secure, most businesses are working to reduce their risk profiles. For example, a company that extends credit can use data analytics to better understand if a customer poses a higher-than-average risk of defaulting. Other companies may use predictive analytics to better understand whether their insurance coverage is adequate.
- **Operational efficiency**: More efficient workflows translate to improved profit margins. For example, understanding when a vehicle in a fleet used for delivery is going to need maintenance before it's broken down on the side of the road means deliveries are made on time, without the additional costs of having the vehicle towed and bringing in another employee to complete the delivery.
- **Improved decision making:** Running any business involves making calculated decisions. Any expansion or addition to a product line or other form of growth requires balancing the inherent risk with the potential outcome. Predictive analytics can provide insight to inform the decision-making process and offer a competitive advantage.

Applications of predictive analytics:

Fraud Detection

Financial services can use predictive analytics to examine transactions, trends, and patterns. If any of this activity appears irregular, an institution can investigate it for fraudulent activity. This may be done by analyzing activity between bank accounts or analyzing when certain transactions occur.

Predictive Analytics vs. Machine Learning

A common misconception is that predictive analytics and machine learning are the same things. Predictive analytics help us understand possible future occurrences by analyzing the past. At its core, predictive analytics includes a series of statistical techniques (including machine learning, predictive modeling, and data mining) and uses statistics (both historical and current) to estimate, or predict, future outcomes.

Credit

Credit scoring makes extensive use of predictive analytics. When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to perform on any credit extended.

Underwriting

Data and predictive analytics play an important role in underwriting. Insurance companies examine policy applicants to determine the likelihood of having to pay out for a future claim based on the current risk pool of similar policyholders, as well as past events that have resulted in payouts. Predictive models that consider characteristics in comparison to data about past policyholders and claims are routinely used by actuaries.

Marketing

Individuals who work in this field look at how consumers have reacted to the overall economy when planning on a new campaign. They can use these shifts in demographics to determine if the current mix of products will entice consumers to make a purchase.

Active traders, meanwhile, look at a variety of metrics based on past events when deciding whether to buy or sell a security. Moving averages, bands, and breakpoints are based on historical data and are used to forecast future price movements.

Supply Chain

Supply chain analytics is used to predict and manage inventory levels and pricing strategies. Supply chain predictive analytics use historical data and statistical models to forecast future supply chain performance, demand, and potential disruptions.

This helps businesses proactively identify and address risks, optimize resources and processes, and improve decision-making. These steps allow companies to forecast what materials will be on hand at any given moment and whether there will be any shortages.

Human Resources

Human resources uses predictive analytics to improve various processes, such as forecasting future workforce needs and skills requirements or analyzing employee data to identify factors that contribute to high turnover rates.

Predictive analytics can also analyze an employee's performance, skills, and preferences to predict their career progression and help with career development planning in addition to forecasting diversity or inclusion initiatives.

Types of Predictive Analytical Models

There are three common techniques used in predictive analytics: Decision trees, neural networks, and regression. Read more about each of these below.

Decision Trees

If you want to understand what leads to someone's decisions, then you may find decision trees useful. This type of model places data into different sections based on certain variables, such as price or market capitalization. Just as the name implies, it looks like a tree with individual branches and leaves. Branches indicate the choices available while individual leaves represent a particular decision.

Decision trees are the simplest models because they're easy to understand and dissect. They're also very useful when you need to make a decision in a short period of time.

Regression

This is the model that is used the most in statistical analysis. Use it when you want to determine patterns in large sets of data and when there's a linear relationship between the inputs. This method works by figuring out a formula, which represents the relationship between all the inputs found in the dataset. For example, you can use regression to figure out how price and other key factors can shape the performance of a security.

Neural Networks

Neural networks were developed as a form of predictive analytics by imitating the way the human brain works. This model can deal with complex data relationships using artificial intelligence and pattern recognition. Use it if you have several hurdles that you need to overcome like when you have too much data on hand, when you don't have the formula you need to help you find a relationship between the inputs and outputs in your dataset, or when you need to make predictions rather than come up with explanations.

Cluster Models

Clustering describes the method of aggregating data that share similar attributes. Consider a large online retailer like Amazon.

Amazon can cluster sales based on the quantity purchased or it can cluster sales based on the average account age of its consumer. By separating data into similar groups based on shared features, analysts may be able to identify other characteristics that define future activity.

Time Series Modeling

Sometimes, data relates to time, and specific predictive analytics rely on the relationship between what happens when. These types of models assess inputs at specific frequencies such as daily, weekly, or monthly iterations. Then, analytical models seek seasonality, trends, or behavioral patterns based on timing. This type of predictive model can be useful to predict when peak customer service periods are needed or when specific sales will be made.

What Is the Best Model for Predictive Analytics?

The best model for predictive analytics depends on several factors, such as the type of data, the objective of the analysis, the complexity of the problem, and the desired accuracy of the results. The best model to choose from may range from linear regression, neural networks, clustering, or decision trees.

- Predictive analytics uses statistics and modeling techniques to determine future performance.
- Industries and disciplines, such as insurance and marketing, use predictive techniques to make important decisions.
- Predictive models help make weather forecasts, develop video games, translate voice-to-text messages, customer service decisions, and develop investment portfolios.
- People often confuse predictive analytics with machine learning even though the two are different disciplines.
- Types of predictive models include decision trees, regression, and neural networks.

Popular predictive analytics models include classification, clustering, and time series models.

- Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behavior to recommend new purchases.

- Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.

- Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.

- Customer insight engines that integrate a wide variety of customer related info, including sentiment, behavior, and even emotions.

- Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.

- Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs— or what potential competitive strategies should be used in a global business that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

Software as a Service BI

- Software-as-a-Service Business Intelligence (SaaS BI) is a business intelligence (BI) delivery model in which applications are implemented outside of a company and usually employed at a hosted location accessed by an end user via protected Internet access.
- SaaS BI generally implies a pay-as-you-go or subscription model, versus the conventional software licensing model with annual maintenance or license fees.
- SaaS BI is also known as cloud BI or on-demand BI.
- SaaS BI allows organizations to use BI tools without on-site installation or maintenance, allowing customers to concentrate on generating analytic queries and BI reports, rather than unnecessary tasks. The SaaS BI approach also allows organizations to broaden their BI systems as usage is increased. Heavy equipment purchases are not required because there are no on-premise deployments.
- SaaS BI can be a fit if there is no available budget to purchase BI software or related hardware. Because there is no upfront purchase expense or extra staffing demands for handling the BI system, the total cost of ownership (TCO) may be much lower than procedures involving traditional on-premise software.

Our first question for James was why his business was so successful: In addition to the Omniture people, several other reasons stand out to me. They include:

- **Scaling the SaaS delivery model.** We built Omniture from the ground up to be SaaS and we understood the math better than the competition. We invented a concept called the Magic Number.The Magic Number helps you look at your SaaS business and helps you understand the value you are creating when standard GAAP accounting numbers would lead you to believe the opposite.

■ **Killer sales organization.** Once we had a few well-known customers like HP, eBay, and Gannett, we stepped on the pedal from a competitive standpoint and really went to battle against the other sales organizations and we won. We focused the whole company on sales.

A focus on customer success. We had 98 percent retention rate. Customer happiness and success were always first because in a SaaS business, unlike traditional enterprise software, it ’s too easy for customers to leave if they are not happy. James explained the three market reasons why he started Domo, knowing we had to fix three problems in traditional BI. Here is a summary in his own words:

1**. Relieving the IT choke point.** Removing the friction for BI to become useful and enabling IT to be more strategic by enabling self-service BI.

2**. Transforming BI from cost center to a revenue generator.** Addresses a very common frustration that I ’ve experienced as a CEO and that other CEOs have shared with me . . . now that we ’ve invested in capturing all this data—how do we benefit from it?

3. **The user experience.** Is where we are putting all our marbles. Today ’s BI is not designed for the end user. It ’s not intuitive, it ’s not accessible, it ’s not real time, and it doesn ’t meet the expectations of today ’s consumers of technology, who expect a much more connected experience than enterprise software delivers.

We ’ll deliver an experience with BI that redefines BI and is unlike anything seen to date.

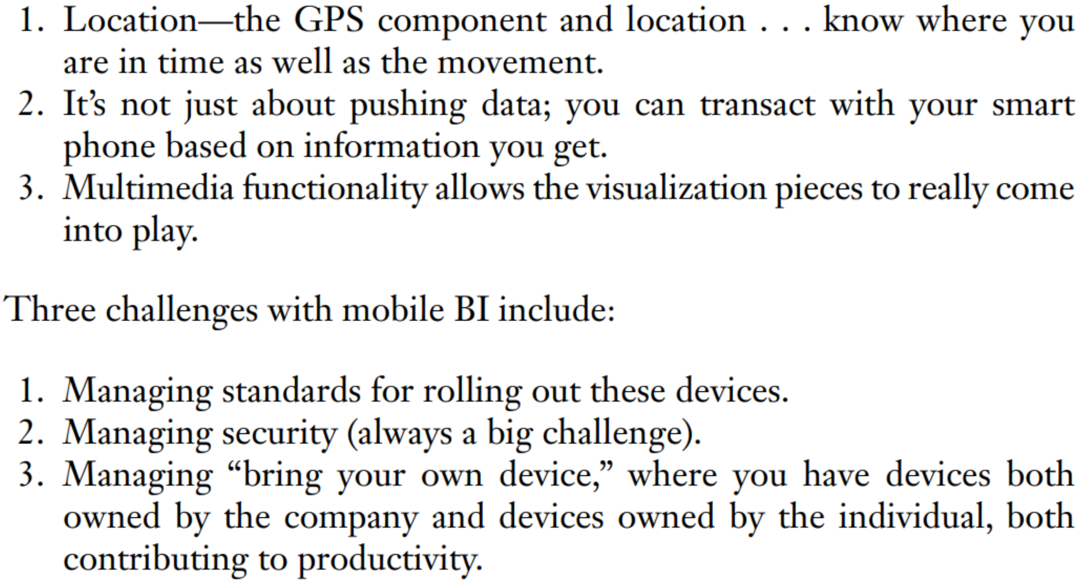
Mobile Business Intelligence Is Going Mainstream

- The definition of mobile BI refers to the access and use of information via mobile devices.

- With the increasing use of mobile devices for business – not only in management positions – mobile BI is able to bring business intelligence and analytics closer to the user when done properly.
- Whether during a train journey, in the airport departure lounge or during a meeting break, information can be consumed almost anywhere and anytime with mobile BI.
- Mobile BI – driven by the success of mobile devices – was considered by many as a big wave in BI and analytics a few years ago. Nowadays, there is a level of disillusion in the market and users attach much less importance to this trend.

Ease of Mobile Application Deployment

Three elements that have impacted the viability of mobile BI:



1. Location—the GPS component and location . . . know where you are in time as well as the movement.
2. It's not just about pushing data; you can transact with your smart phone based on information you get.
3. Multimedia functionality allows the visualization pieces to really come into play.

Three challenges with mobile BI include:

1. Managing standards for rolling out these devices.
2. Managing security (always a big challenge).
3. Managing "bring your own device," where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

Crowdsourcing Analytics

What is crowdsourcing in data analytics?

What Is Crowdsourcing? Crowdsourcing involves **obtaining work, information, or opinions from a large group of people who submit their data via the Internet, social media, and smartphone apps**. People involved in crowdsourcing sometimes work as paid freelancers, while others perform small tasks voluntarily.

In October 2006, Netflix, an online DVD rental business, announced a contest to create a new predictive model for recommending movies based on past user ratings.

- Netflix already had an algorithm to solve the problem but thought there was an opportunity to realize additional model "lift," which would translate to huge top-line revenue.

- Netflix was an innovator in a space now being termed crowdsourcing. Crowdsourcing is a recognition that you can ’t possibly always have the best and brightest internal people to solve all your big problems.

- By creating an open, competitive environment with clear rules and goals, Netflix realized their objective and, yes, they did create a lot of buzz about their organization in the process.

- Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models.

- Kaggle describes itself as “an innovative solution for statistical/analytics outsourcing.” That ’s a very formal way of saying that Kaggle manages competitions among the world ’s best data scientists.

How it works?

- Corporations, governments, and research laboratories are confronted with complex statistical challenges.
- They describe the problems to Kaggle and provide data sets.
- Kaggle converts the problems and the data into contests that are posted on its web site.
- The contests feature cash prizes ranging in value from $100 to $3 million.
- Kaggle ’s clients range in size from tiny start-ups to multinational corporations such as Ford Motor Company and government agencies such as NASA.

Inter- and Trans-Firewall Analytics

In the health care industry, rich consumer insights can be generated by collaborating on data and insights from the health insurance provider, pharmacy delivering the drugs, and the drug manufacturer.

In-fact, this is not necessarily limited to companies within the traditional demand-supply value chain.

For example, there are instances where a retailer and a social media company can come together to share insights on consumer behavior that will benefit both players.

Some of the more progressive companies are taking this a step further and working on leveraging the large volumes of data outside the firewall such as social data, location data, and so forth.

In other words, it will be not very long before internal data and insights from within the firewall is no longer a differentiator. We see this trend as the move from intra- to inter- and trans-firewall analytics.

Today they are doing intra-firewall analytics with data within the firewall. Tomorrow they will be collaborating on insights with other companies to do inter-firewall analytics as well as leveraging the public domain spaces to do trans-firewall analytics.

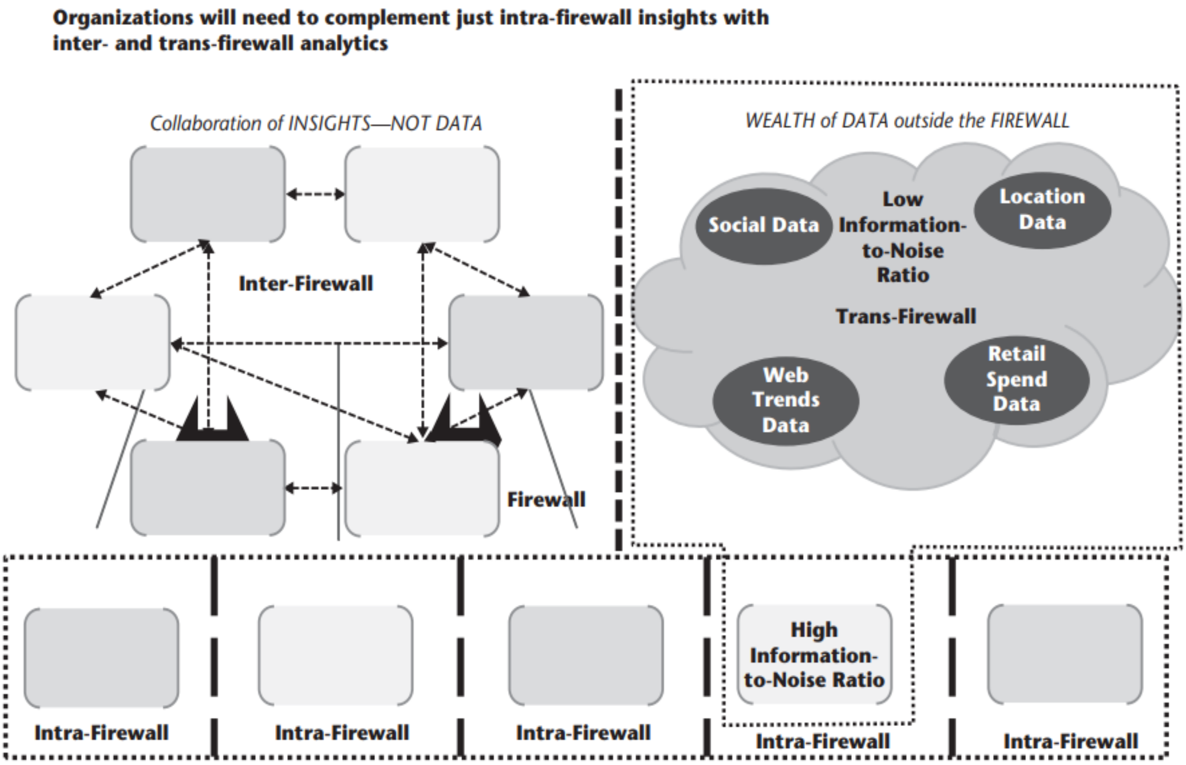
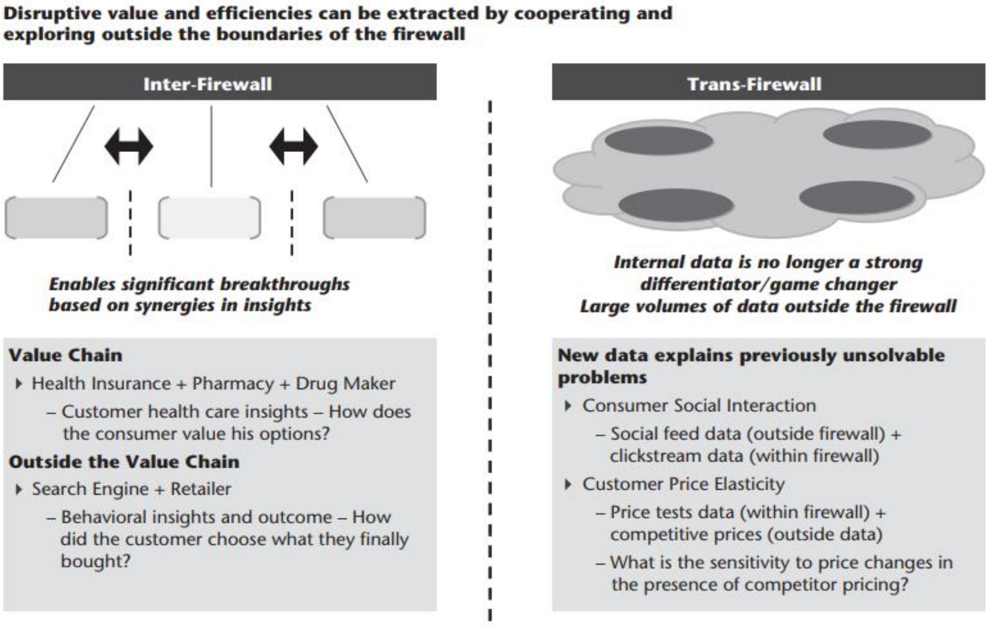

Organizations will need to complement just intra-firewall insights with inter- and trans-firewall analytics
Collaboration of INSIGHTS—NOT DATA
Inter-Firewall
Firewall
WEALTH of DATA outside the FIREWALL
Social Data
Low Information-to-Noise Ratio
Location Data
Trans-Firewall
Web Trends Data
Retail Spend Data
High Information-to-Noise Ratio
Intra-Firewall
Intra-Firewall
Intra-Firewall
Intra-Firewall
Intra-Firewall

Figure 3.1 Inter- and Trans-Firewall Analytics
Source: Mu Sigma and author Ambiga is a cofounder.



Disruptive value and efficiencies can be extracted by cooperating and exploring outside the boundaries of the firewall

Inter-Firewall

Enables significant breakthroughs based on synergies in insights

Value Chain

- Health Insurance + Pharmacy + Drug Maker
  - Customer health care insights – How does the consumer value his options?

Outside the Value Chain

- Search Engine + Retailer
  - Behavioral insights and outcome – How did the customer choose what they finally bought?

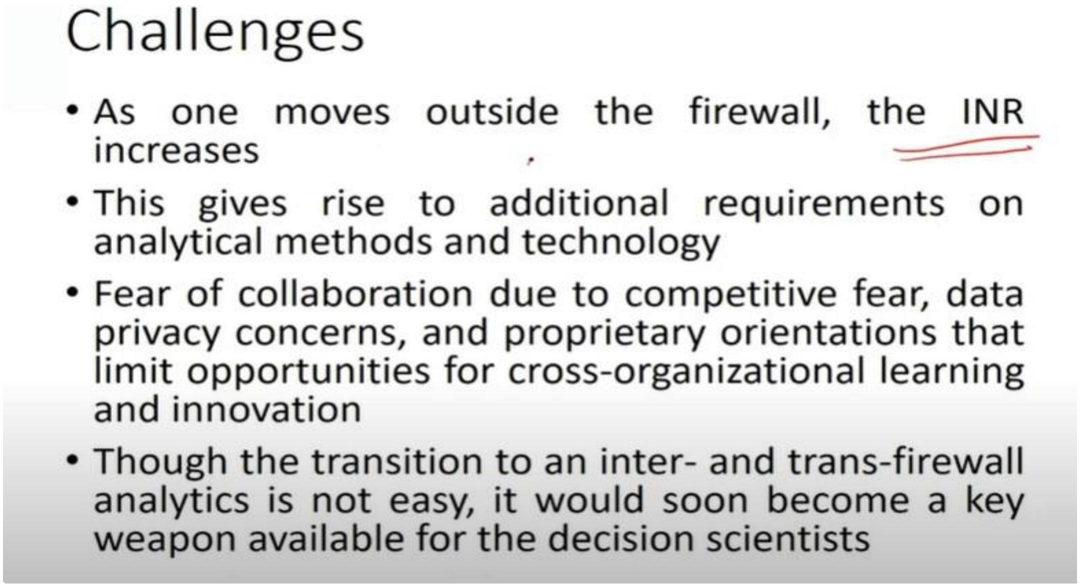
Trans-Firewall

Internal data is no longer a strong differentiator/game changer
Large volumes of data outside the firewall

New data explains previously unsolvable problems

- Consumer Social Interaction
  - Social feed data (outside firewall) + clickstream data (within firewall)
- Customer Price Elasticity
  - Price tests data (within firewall) + competitive prices (outside data)
  - What is the sensitivity to price changes in the presence of competitor pricing?

Figure 3.2 Value Chain for Inter-Firewall and Trans-Firewall Analytics
Source: Mu Sigma.



Challenges

- As one moves outside the firewall, the INR increases
- This gives rise to additional requirements on analytical methods and technology
- Fear of collaboration due to competitive fear, data privacy concerns, and proprietary orientations that limit opportunities for cross-organizational learning and innovation
- Though the transition to an inter- and trans-firewall analytics is not easy, it would soon become a key weapon available for the decision scientists