Big Data Spectrum

Infosys®

POWERED BY INTELLECT
DRIVEN BY VALUES

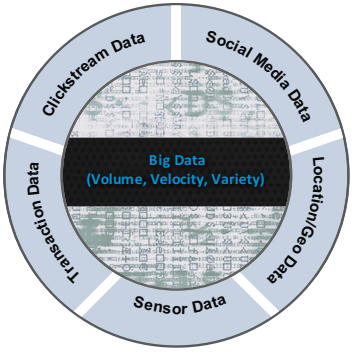
Contents

Introduction

What is Big Data?

Today we live in the digital world. With increased digitization the amount of structured and unstructured data being created and stored is exploding. The data is being generated from various sources - transactions, social media, sensors, digital images, videos, audios and clickstreams for domains including healthcare, retail, energy and utilities. In addition to business and organizations, individuals contribute to the data volume. For instance, 30 billion content are being shared on Facebook every month; the photos viewed every 16 seconds in Picasa could cover a football field.

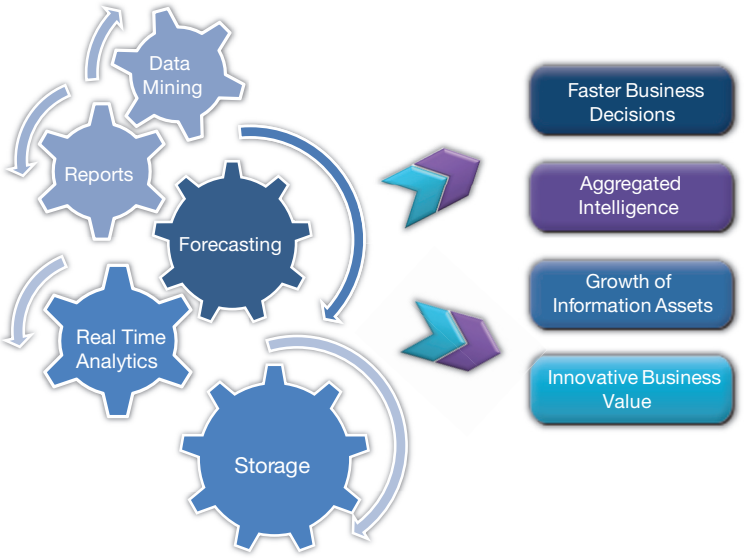
It gets more interesting. IDC terms this as the 'Digital Universe' and predicts that this digital universe is set to explode to an unimaginable 8 Zeta bytes by the year 2015. This would roughly be a stack of DVD's from Earth all the way to Mars. The term "Big Data" was coined to address this massive volume of data storage and processing.



It is increasingly becoming imperative for organizations to mine this data to stay Competitive. Analyzing data can provide significant competitive advantage for an enterprise. The data when analyzed properly leads to a wealth of information which helps the businesses to redefine strategies. However the current volume of big data sets are too complicated to be managed and processed by conventional relational databases & data warehousing technologies.

The volume, variety and velocity of Big Data causes performance problems when being created, managed and analyzed using the conventional data processing techniques. Using conventional techniques for Big Data storage and analysis is less efficient as memory access is slower. The data collection is also challenging as the volume and variety of data has to be derived from sources of different types. The other major challenge in using the existing techniques is they require high end hardware to handle the data with a huge volume, velocity and variety.

Big Data is a relatively new phenomenon. As with any new adoption, the adoption of Big Data depends on the tangible benefits it provides to Business. Large data sets which are considered as information overload are invariably treasure troves for business insights. The volume of data sets has immense value that can improve the business forecast, help in decision making, deciding business strategies over the competitors. For instance, Facebook, blogs and twitter data gives insights on current business trends.



The data sets are beyond the capability of humans to analyze manually. Big data tools have the ability to run ad-hoc queries against the large data sets in less time with a reasonable performance. For instance, in retail domain understanding what makes a buyer to look into a product online, sentiment analysis of a product based on the Facebook, tweet and blogs are of great value to the business. This will enable the business to improve their services for customers.

Big Data analysis enables the executives to get the relevant data in less time for making decisions. Big Data can pave way for fraudulent analysis, customer segmentation based on the store behavior analysis, loyalty programs that identifies and targets the customers. This enables us to perform innovative analysis which indeed changes the way we think about data.

Exploring Big Data Spectrum

With unstructured data dominating the world of data, the way to exploit it is just becoming clearer. Information proliferation is playing a vital role in leveraging the opportunities and is also presenting a plethora of challenges.

The industry opportunities presented by the plethora of data are plenty. To understand how to leverage Big Data opportunities is a clear need to the business.Big Data spectrum covers use case from five different industries Retail, Airlines, Auto, Financial Services and Energy.

All opportunities come with a set of challenges. The way to know and address these challenges is discussed in the Key Challenges section. To name a few: Data Privacy,

Data Security, Integrating various technologies, catering to real time flow of data and leveraging cloud computing.

To dive deep into Big Data technology with the goal of having a quick, managed and quality implementation, a set of enablers were designed by the Architects at Infosys. The section on Adoption Enablers gives the insights into these enablers.

The sections are interleaved with the viewpoints from Phil Shelly, CTO, Sears Holdings Corporation, Doug Cutting, Co-founder of Apache Hadoop (popularly known as father of Big Data) and Kris Gopalakrishnan, Co-Chairman, Infosys Ltd.

Q&A

Phil Shelley
CTO, Sears Holdings Corporation

Dr. Shelley is a member of CIO forum, Big Data Chicago forum

Phil, Sears is one of early adopter of Big Data. What are the sweet spot use cases in the retail industry?

Transactional data such as POS, Web-based activity, loyalty-based activity, product push, seasonality, weather patterns and major trends that affect retail, business value that can be mined from this data. In addition to this you add the data in the social space and the sheer amount of this data is way beyond what traditional database solutions can handle. That is where Hadoop plays a role, to capture and keep this data in the finest level of detail.

What kind of challenges are you facing in implementing Big Data solutions?

Hadoop is relatively low cost to implement. However, to get started, you still need some kind of business case. It is a good idea to start small and have a very specific use case in mind. At the same time, a picture of where big data will be valuable long term is important as well. Focusing on a key use case that can demonstrate business value is probably the way to start, for any company a big-bang approach is not something I would recommend.

How Big Data management integrates with your Enterprise systems?

Hadoop is not a panacea. Big data solutions will be a hybrid of traditional databases, data warehouse appliances and Hadoop. The combination of high-speed SQL access and the heavy lifting of Hadoop can work together very well. This means that you need to synchronize the data between these data sources. One way to do this is to aggregate your data before you send it to an appliance. What data needs to be shared and how the synchronization happens will need some careful thinking. At least for the next few years it's going to be an ecosystem of Hadoop combined with a more traditional database system.

How to handle Data Privacy and Security issues in the Big Data management?

Personally I would not put very sensitive data on a public cloud because the risk of exposure could be catastrophic to the company. A private cloud which is co-located with my data center or, a virtual private cloud that is physically caged, are approaches I would recommend. Out of the box Hadoop has security limitations. You have to explicitly design your data for security. There are ways of securing credit card and

personal information. I would recommend that anyone looking to secure such data look to some help on how to structure a big data solution and not expose themselves. This is an area that is somewhat new and prone to lax security.

Phil, One last question, as a CTO of Sears how do you see the connection between real-time Digital Enterprise and Big Data?

Today Hadoop can have near real time copies of transactional data and near real time batch reporting with as little as minutes of latency. You can process the data near real time and then access from a data mart in real time, which create many possibilities. But it is really going to be an ecosystem of the right tools for the right jobs.

Making it Real – Industry Use Cases

Retail – Extreme Personalization

Use case Context

There was a complete directory of the whole World Web in the beginning. One knew about all the servers that were in existence. Later other web directories appeared ex: Yahoo, AltaVista etc. which kept a hierarchy of the web pages based on their topics. In 1998 Google changed everything. Crawling, Indexing and Searching paved the way for Personalization and Clickstream analysis, by which personalized recommendations based on social, geographical and navigation could be made. This personal flavor gives higher value and causes customers to become more loyal and more profitable for businesses. Now, the channels by which voice of customer can be collected have multiplied leading to the big data explosion, and retailers have a huge opportunity to present customers with even more personalized promotions, deals and recommendations.

Challenges

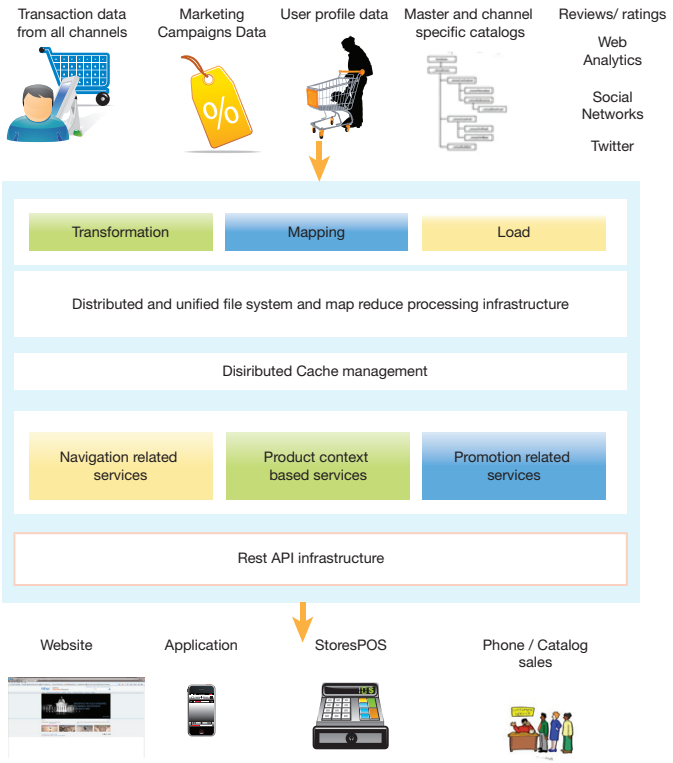
One of the significant challenges in architecting such a personalization system is the amount and diversity of data that has to be handled. For example, websites today generate user activity data that could easily run into Terabytes in a matter of months. Equally problematic is the different formats and system interfaces. Once the data is loaded, the system applies correlation techniques to correlate the data and draw inferences about the preferences of individual customers. The traditional relational data warehousing OLAP based systems struggle to process this massive amount of high velocity data and provide insights. There is a high latency between the user's shopping activity and the generation of recommendation as well as limited granularity which decreases the relevance of the recommendation to the end customer.

Solution and Architecture

The personalization System sources information about the customer profile, orders, preferences and opinion from multiple sources – some within the enterprise and some outside. Examples of sources are the enterprise order management system (or channel specific order management systems where applicable), customer browse analytics data from the website, opinions and reviews from social networks, etc. In addition to customer data, some master data like the sales catalog and marketing campaigns are loaded.

There are 3 basic levels of inferences that can be done about what a consumer might be interested in:

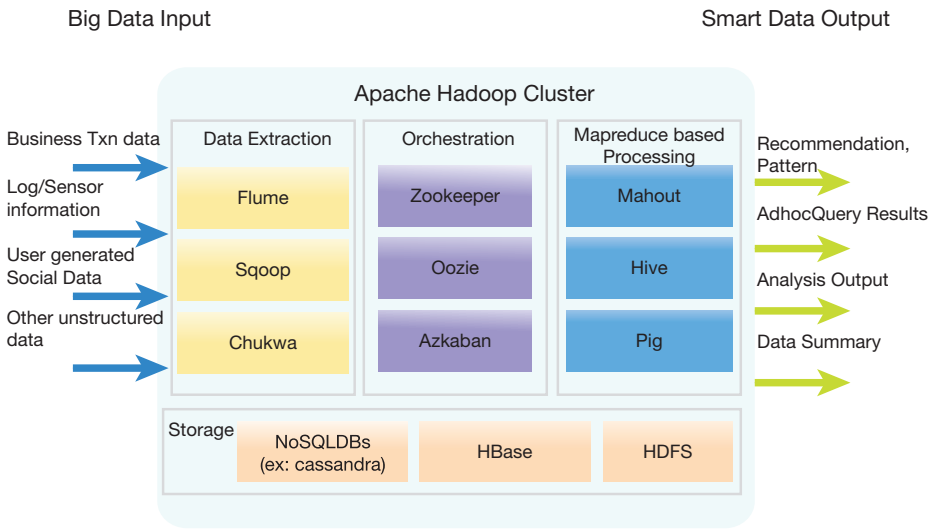
1. Based on past actions and opinions.
2. Based on similar users (actual friends as defined by the consumer in social networks, statistically derived segments based on behavior).
3. Based on general public behavior



Inferences from the personalization system can be consumed by multiple scenarios and systems. A website can use that to show content that is more relevant to the customer. The store systems can provide more relevant suggestions and customer service representatives might know more about the consumers than the consumers themselves!

A personalization system has to process the data from various sources mentioned above and come up with personalized recommendations and business intelligence for individual customers. This is not a simple task. This involves processing tera/peta bytes of data obtained from various sources. Traditional systems are not efficient and scalable in storing and processing this much amount of data. Big data processing solutions are the need for such a system.

Typical Architecture of a Big Data processing system using Apache Hadoop Stack is shown below. In the context of extreme personalization, we can use the data extraction tools to load the data into a Hadoop system and use Mahout which is a scalable machine learning library which comes with a lot of algorithms for pattern mining, collaborative filtering and recommendations. The latest trend is the evolution of real-time systems like HStreaming and Twitter Storm which perform the analysis in an almost instant fashion compared to Hadoop which is run in batch mode.



Business Value

Real-time Personalized insights: By combining inputs from various channels (social, location, history etc.) and analyzing them real-time, customers can be presented with almost instant recommendations. For example, if a customer tweets "I like Xbox" , the system can provide recommendations related to Xbox when she logs into the ecommerce site or as an ad on her social network profile or even send an Xbox promotion coupon to her mobile if she's shopping in-store. This kind of highly personalized and instant recommendation is being experimented and will become more prevalent going forward.

Personalized Presentation Filtering: One of the fundamental things that can be offered is the ability to present content which is tuned to the preferences of the consumer. This could be in terms of product types (Wii related vs. Xbox related or full sleeve vs. half sleeve), brands (Sony vs. Samsung) or Prices (costliest vs. cheapest) or something else that we know about her. This can be provided as filtered navigation in a website or as a suggestive selling tip to a customer service representative while they are speaking to the consumer.

Context-specific and Personalized External Content Aggregation: Presentation of context-specific information that makes sense for the consumer is a key capability. A good example is the relevance of Social context. If we are showing a product and can show the consumer that out of the 1500 people who said that they liked the product 15 are friends (with the capability to know who those 15 were as well), the impact and relevance of that would be significant. This service is relevant only for electronic channels.

Personalized Promotional Content: Different consumers get attracted by different value propositions. Some like direct price cuts, some like more for the same money

(please note that they are exactly not the same), while some others believe in getting more loyalty points. Showing the most appropriate promotion/offer based on their interests is another important capability that a personalized system can provide.

Big data processing solutions can process vast amount of data (tera/peta bytes) from various sources like browsing history, social network behavior, brand loyality, general public opinion about the product obtained from various social networks etc. This kind of extremely useful and tailor-made information can only be obtained using Big Data processing solutions and retailers must leverage them to make their business more appealing and personal to customers.

Airlines – Smart Pricing

Use Case Context

Air transportation is one of the toughest and most dynamic industries in the world. Constantly troubled by factors like oil prices, thin profit margins, environmental concerns, employee agitations etc, it is truly a world of survival of the fittest. Understanding the end consumer needs and discovering the competitive and attractive price to consumer is always a challenge. Airlines strive to achieve key differentiators like providing a personalized experience, enhance the brand, predictive intelligent systems for identifying profit and loss factors etc. In the airline industry, a few high-value customers can generate much more revenue than a number of low-value customers, which highlights the importance of CRM systems. Some of the questions faced by the airlines in relation to CRM include:

- How are the customers segmented? What share of profits does each customer segment bring in?
- Some customers deserve greater attention than others. How to identify them from the frequent flyers?
- What tactics should be adopted to acquire, convert, retain and engage customers?

Personalization and Fare processing involves large data sets. While the current systems leverage conventional enterprise information, new data sources such as social media, web logs, call center logs & competitor pricings would have to be considered for personalization requirements.

Here the challenges include highly scattered data sources, huge effort for data integration, and high cost of data warehousing and storage solutions.

Scenario 1: Fare Processing:

Context and Challenges

Airlines classify fares into the main classes of service (Economy, Business, First etc.), and each of these are subdivided into booking classes. The number of booking classes

depends on factors like aircraft type, sector of the flight, flight date etc. The objective here is to have a greater level of control over the type of fares sold. When making a fare change, the fare is re-priced, which means recalculating each fare, taking into consideration factors like:

- Flown data / passenger booking information
- Forecasted data - load factor information
- Inventory
- Currency exchange rate at different points of sale

Information about these factors will be distributed across various databases. Naturally, an attempt to process these data will run into all the above mentioned problems. And these data sources may not be sufficient to discover a price aligned to the customer's point of view.

Opportunity – Big Data for Fare Processing:

Fare processing involves extracting and consolidating information from external sources such as Agent systems, External Pricing systems, and internal systems such as Revenue Accounting, Forecasting, Inventory and Yield Management systems. This data itself will run into tens of terabytes. Newer data sources such as Social Media conversations about pricing decisions & competitors, and un-structured data in enterprise like customer service e-mail data, call center logs etc. will push the data volumes even further.

In this context, the high-cost data warehousing solutions are confronted with the following challenges.

- Expensive hardware and software : Cost grows with data size
- Analytics process needs to be 100% customizable, and not be governed by the confines of SQL
- Analytics process should not get extended by the size of the data; terabytes should be processed in minutes rather than days or hours
- Data loss is unacceptable, the solution requires high availability and failover
- Learning curve should not be steep

Scenarios 2: Personalization for high-value customers

Context and Challenges

The airline industry is one where a few high value customers are more significant than many low value ones. Naturally, the winners are those who can do predictive analytics on the data and provide a personalized flying experience. A wealth of customer behavioral data can be gleaned from websites – like route and class preferences, frequency of flying, baggage, dining information, geo-location etc. - to name a few.

This type of analytics requires accumulation of data from multiple sources, frequent processing of high volume data, flexibility and agility in the processing logic. These are all pain areas for traditional data warehouse solutions and these get compounded as the data volume grows.

Opportunity – Holistic data analytics and newer data

Analyzing the customer's data from various systems including Loyalty, CRM, and Sales & Marketing etc. and data from partners systems along with data from social media based on customer social profile can help airlines to create deeper personalization preference of customers and also understand their current social status and preferences.

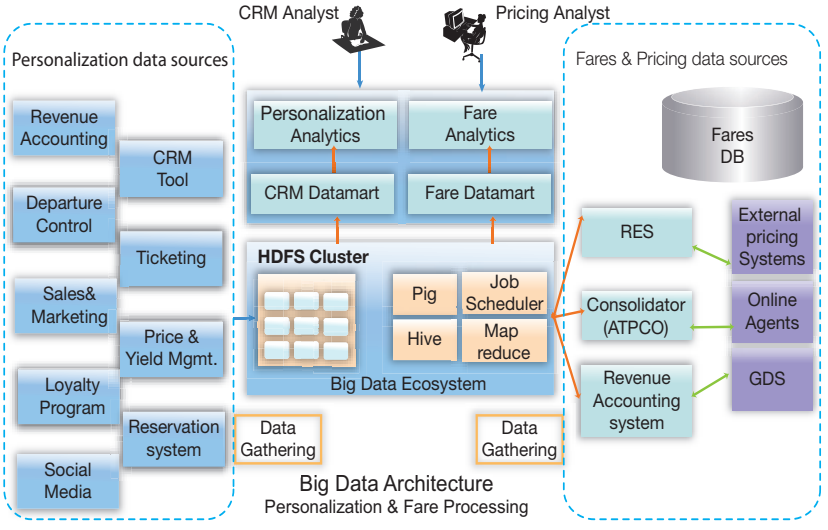
Solution and Architecture

Big Data Architecture leveraging Hadoop stack provides ability to extract, aggregate, load and process large volume of data in a distributed manner which in turn reduce the complexity and overall turn-around time in processing large volume of data.

Proposed solution is aimed at primarily addressing the following challenges:

- Ability to manage and pre-process large volume of enterprise data, partners data and logs data and transform these data into meaningful data for analytics (Analysis, Preprocessing for Fare calculation and personalization services)
- Run analytics on un-structured data from social and other sources to derive newer dimension such as sentiment, buzz-words, root causes from customer interactions and user created content

Data extracts from various feeder systems are copied to the Big Data file system. The Analytics platform will trigger the processing of this data, while responding to



information needs for various analytical scenarios. The Analytics platforms contain the algorithms tailored to the business process, to extract/mine meaningful information out of the raw data.

Business Value

By switching over to a Big Data based solution, it is estimated that the fare processing times can be cut down from days to a matter of hours or even minutes. The airline industry operates in a high risk domain vulnerable to a large number of factors, and this agility will be a vital tool in maximizing revenues. Big data architecture solutions for personalization helps in understanding and deriving granular personalization parameters and understand customers social status and needs based on their interaction in social media systems.

The potential benefits for an airline could include:-

- Increased revenue & profit
  - better fare management
  - efficient mechanisms to handle fare changes
- Faster decision making and reduced time-to-market for key fares
- Introduction of competitive fares & reactive fare response
- Increased efficiencies to the airline in managing fare strategy for different sales channels
- Deeper understanding of customers and personalization (N=1)

Auto – Warranty & Insurance Efficiency

Use case context

Two of the biggest threats to the auto insurance companies are, economic downturn, which is making money unavailable, and other one is, insurance frauds. An insurance claim starts with the process of applying for a policy and setting up the right premium and then finishes with identifying the valid claims and minimize the frauds. The investigation done during the premium setting can help in minimizing the losses which can occur during the claim settlement to a great extent.

These days there are lot of tools available in the market to help the companies mine the data and automate this whole process. These tools not only streamline the process but helps by providing suggestion in various steps. However, the information availability to the companies is huge and growing everyday in leaps and bounds. With the recent explosion of the information from both external channels such as the social media sites and internal channels such as BPOs, Collaboration mediums etc, it has become really challenging to mine such huge data to get the meaningful insight. But at the same time, ability of doing the same provides any insurance company a huge competitive advantage.

Key challenges with regards to typical Auto Insurance workflow can be:

- Verifying the data collected from the customer
- Profile customer behavior, and social networking influence
- Identifying the right insurance premium amount
- Verifying the claims raised
- Fraud detections by analyzing data from the disparate systems
- Exact claim reimbursement

Solution and Architecture

The Big Data approach brings in a new dimension in solution to the challenges mentioned above. With Big Data the approach in each stage mentioned above can be augmented by verifying the data collected from multiple internal and external sources and running artificial intelligence across all those information to identify any pattern of frauds or any other possible compromise. With Big Data technologies being in place, solution for two core challenges integral to these stages become practical and affordable –

1. Ability to store and crunch any volume of data in a cost effective way
2. Ability to model statistically a rare event like fraud which needs sample data size close to the entire population to capture and predict right type of signatures of the rare events

The potential data sources which can be leveraged for the same being

- Source: Internal Systems –
    - Collect the data about the customer from the Internal CRM system.
    - Check the other products sold to the customer from In-house ERP, CRM systems.
    - Check the credit history and based on that provide the necessary quotes from other internal systems.
- Source: External Systems –
    - Collect the data from the social networking sites regarding the behavior of the customer – profile customers on their behavior, sentiments, social net-worth, usage patterns, click stream analysis for their likes and dislikes, behavior and spending patterns.
    - Check the reviews and ratings of the product for which the insurance is required. For example if the insurance is required for the particular brand of car then reviews about the cars can be mined and by using sentiment analysis, the cars can be put in different categories. The cars with better reviews and ratings can be offered with lower premium and cars which have bad reviews and which brake down frequently can be offered with higher insurance premium.

- Check the social network of the user. This is can help in identifying the status of the customer and even will help in identifying potential customers.
- Insurance companies can even make use of the data such as the customers interest in the car racing or his network which is quite active in rally racing etc.
- Some of the other data which can be obtained from the external systems includes the data to check the credit worthiness of the customer from the external 3rd party rating sites

The solution architecture breaks the entire landscape into Acquisition, Integration and Information Delivery/Insights layers. Those three layers provide a seamless integration, and flexible options to plug-n-play additional sources and delivery channels hiding the underlying data complexity in each layer. The core to managing such architecture is towards handling the large volume of information, and the varied formats of data flowing into the system. Hence the solution's centered on big data and data virtualization techniques to manage this burst of information.

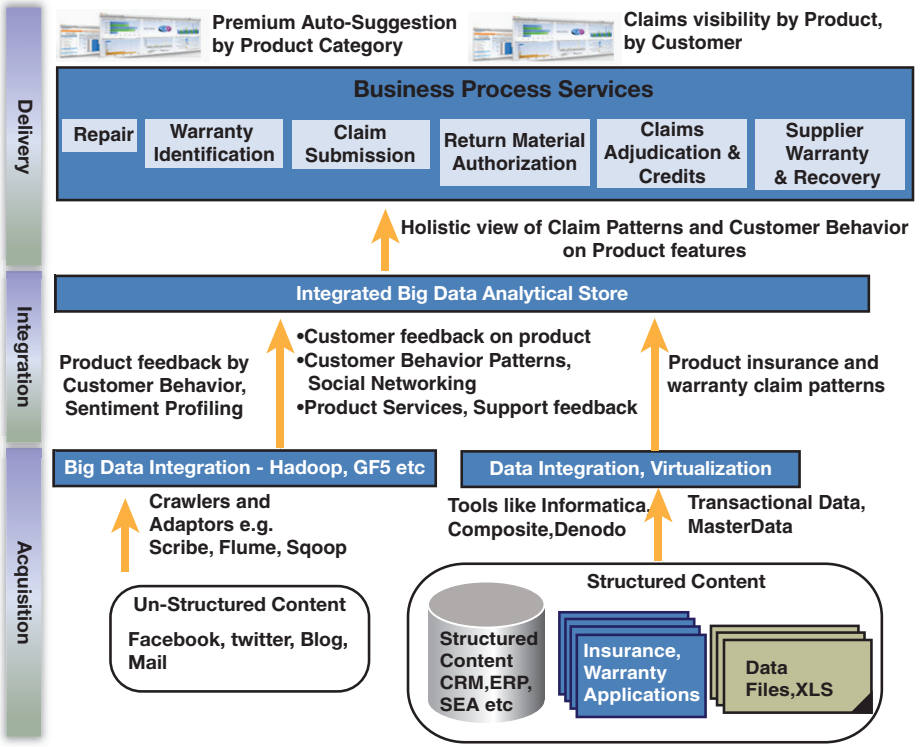
Acquisition Layer – Once the sources of both structured and un-structured data are identified, both externally and internally the right set of adapters and crawlers are required to extract the Big data from public/private locations outside the boundaries of organization. This is most critical aspect, and the right set of business rules on what needs to be pulled, requires an upfront preparation. Idea is to start with smaller number of sources which have more likelihood of having relevant set of data for desired outcome, and later extend this strategy to other sources. In parallel, is to leverage data integration and data virtualization technologies to integrate the relevant data sources which can provide auto insurance policy, customer details, product categories and characteristics, claims history. Tools like Informatica, IBM DataStage, ODI etc for data integration, and Denodo, Composite etc for data virtualization can be leveraged.

Few parameters need to be factored in for better performance:

- Frequency of data capture from transactional systems, and external sources
- Delta feeds, trickle feeds and staging strategies can play an important role
- Effective business rules, technical rules filters in place to reject unwarranted data sets

Integration – This is where post extraction and filtering, the information gets consolidated under a common data model. Integration requires a strong set of mapping rules to map both structured and un-structured (transformed into structured format). The data model should support integrating following set of information:

- Auto insurance policy details by vehicle type
- Claim history by vehicle type and customer
- Customer feedback on product features – sentiments, opinions



- Customer behavior pattern analysis – liking for red color vehicles, sports vehicles etc
- Customer social network analysis – influence, and near neighbor financial capacity etc

Delivery – The final stage of delivering the insights generated from the integrated information as dashboards, scorecards, charts and reports with flexibility for business analysts to explore the details, and correlate the information sets for taking decisions on setting insurance premiums by vehicle and type of customers. The advanced analytics techniques in delivering such information will also help figuring out claim frauds, and unregulated auto insurance policies, claims. The delivery channels can be desktops via portals, mobile devices, internet based application portals etc.

Business Value

With Big Data technologies being in place solution for the two core challenges integral to the problem statement become practical and affordable –

1. Firstly, the ability to store and crunch any volume of data in a cost effective way
2. Secondly, the ability to model statistically a rare event like fraud which needs sample data size close to the entire population to capture and predict right type of signatures of the rare events

And with that the immediate business values this solution can bring, can be categorized in –

- Ability to tag right pricing on Insurance Premium with holistic view and better insight
- Better claims visibility and attribution to product insurance premiums/warranty
- Lesser fraud case and loss

While this solution addresses the specific use case of auto insurance premium advisor, the natural extension of this solution and framework can certainly be applicable to Manufacturing parts warranty, other Insurance premium and claims, fraud detection processes covering domains like Manufacturing, General Insurance and Retail.

Financial Services - Fraud Detection

Use case context

Fraud detection use cases are diverse and their complexity can be characterized by the fact that they are either real-time/batch, use unstructured data or only structured data, use a rules engine or derive a pattern (where the rule is not known but one is looking for any sequence of seemingly unrelated events which may be interpreted as a fraud).

Here we focus on one such use case for elaboration and we will detail out the applicable solution and its architecture.

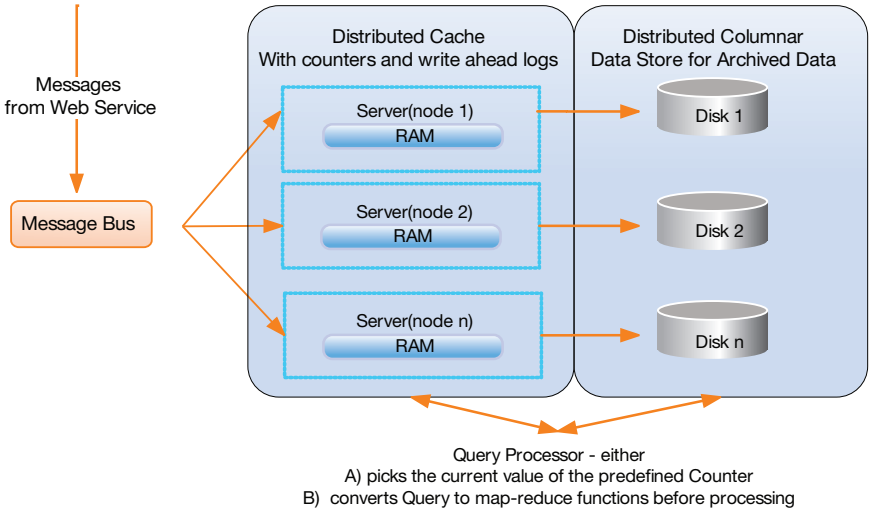
Let's take an example of an online account opening form, which most of the financial service firms provide as an online service. The first step is to validate the identity of the user as entered in the form. Such forms can be misused by hackers to extract a valid combination of name, address, email id, SSNID etc using brute force. Most of these applications would be sending a request to a web-service to check the validity of this prospective client – and usually this web service is provided by an external vendor specializing in credit ratings. The requirement for the credit rating agency would be to intercept calls made to the web-service, look at the patterns of such calls and identify any fraudulent behavior.

At the web-service end the characteristics of the challenges can be stated as follows

- The evaluation of individual request has to be done within seconds (depending on the SLA)
- Fraud evaluation would require historical data related to each request
- Evaluation would be based on a set of predefined rules- which may focus on all the request sent in the past predefined time window

Solution and Architecture

Based on the definition of the problem and its characteristics, we would have the following technical requirements:



1. The System should be able to ingest requests at the rate they come in which will vary over a period of time
2. The latency of the system has to be below specified SLA limits which may not allow the system to store the incoming data before the response is to be evaluated
3. The need is to evaluate a set of incoming response which will need a good amount of RAM to accommodate the data in memory and any derived structures created to store the aggregated over history and current stream
4. With increased load the system may be required to use multiple nodes to ingest the data while keeping the overall validity of the counters across the nodes.
    a. The counters will be needed to be stored in a shared memory pool across the nodes.
    b. The counters will help reduce latency as they would be updated before the data is even written to the disk (for history updates)
5. Distributing the system over multiple nodes will provide the solution with parallelism and also ability to develop a fault-tolerant solution
    a. The distributed nodes as stated in the last point will be able to handle parallel writes across multiple nodes in a peer-to-peer architecture (not in a master –slave architecture)

b. With increased number of nodes the probability of a node failure increases and hence replication of Historical data would be needed across multiple nodes. Similarly the data could be sharded across the nodes to help in parallelizing reads.

Assuming, that the web service takes care of writing a message to the message queue for each request received with appropriate details, a Typical Architecture for such a solution will be composed of the following components

6. The Acquisition layer
    a. This component will read Messages from the message queue and distribute it to a set of worker processes which will continue with the rest of the acquisition process.
    b. Each worker process will be able to Look up the cache for an appropriate data structure based on the message details – if found update the counter. Else create a new one.
7. The distributed cache - The role of the distributed cache will be to act as initial data store based on which the analysis could be done. Thus helping reduce latency between the message arrival and its impact on the measurement. This will need
    a. Initialization of the distributed cache while startup and also on a regular basis while data is flushed to the data disks
    b. Ability to Flush the data in cache to the data disk when the cache size reaches a certain water mark
    c. Ability to create local structure on the node where the message is received and replicate it to the copies situated on other nodes.
    d. Ability to create and maintain a predefined set of replicas of the data structure across the nodes to support fault tolerance
8. The Storage/retrieval layer
    a. Ability to store serialized data structures for the related processing nodes with adequate copies across multiple nodes to handle the fault tolerance in the data storage layer
    b. Ability to provide secondary index on the Data structures for alternate Queries The historical data stored will be time series in nature and columnar distributed data stores would be an appropriate way to handle this.
    c. The Data could be sharded across data nodes to increase the read response.

Business Value

Above mentioned solution provides opportunity to

- Reduce costs – with ability to handle large volumes of varying load using commodity hardware

- Meet Risk requirements – this kind of latency would not be possible in a traditional RDBMS where data would have to be stored and indexed before querying
- Further alerts/event processing can be configured in the CEP to take appropriate action on detection of a fraudulent request.

Energy – Tapping Intelligence in Smart Grid / Meters

Use case context

The two main issues that utility majors face across the world are environmental concerns and Power delivery limitations and disturbances. To address these issues and taking into consideration the technology advancement, electric power grids are being upgraded with smart meters installed at consumers, and other Grid sensors for efficient monitoring of the utility infrastructure. However, the true value of smart grids is unlocked only when the veritable explosion of data is ingested, processed, analyzed and translated into meaningful decisions such as; ability to forecast electricity demand, respond to peak load events, and improve sustainable use of energy by consumers.

The major challenges the utility providers face today are to

- Curb inadequacies in generation, transmission and distribution and inefficient use of electricity
- Reduce technical and commercial losses (AT&C) that lead to substantial energy shortage
- Improve quality of power supply
- Increase revenue collection
- Provide adequate electricity to every household & improve consumer satisfaction

One solution we can look at to address the above challenges is through the implementation of smart grids with Big Data analytics Platform. Smart Grid is a real time, automated system for managing energy demands and responses for optimal efficiency. In a smart grid environment, demand response (DR) optimization is a two-step process consisting of peak demand forecasting and selecting an effective response to it. Both these tasks can greatly benefit from the availability of accurate and real time information on the actual energy use and supplementary factors that affect energy use. Analytical tools can process the consumption of data coming in from an array of smart meters and provide intelligent data which can help the utility company to plan better for capital expenditures. Hence, the software platform that collects, manages and analyzes the information plays a vital role.

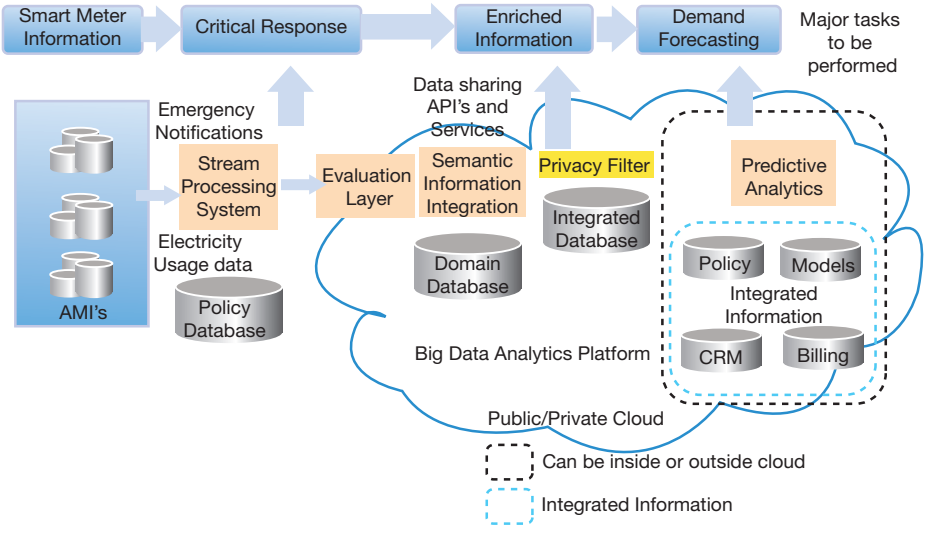
Solution and Architecture

Datasets (such as structured, unstructured and semi-structured data)coming out of smart meters are enormous and amounts to petabytes of data and processing such data in relational DBMS demands lots of investment in terms of cost and time. Big Data technologies come in handy in storing such huge amounts of data. Big Data Analytics Platform (BDAP) can help in analyzing these datasets and provide meaningful information which can be used by the utility major while taking instantaneous as well as short term to long term decisions. Thus BDAP is one such solution to assuage power industry's pain point.

Efficient Information integration and data mining contributes to an architecture that can effectively address this need. The primary tasks are

- Ingest information coming from smart meters
- Detect critical anomalies
- Proceed with non-critical task of annotating smart meter data with domain ontologies (The collective set of information models used by the electricity industry can be viewed as a federation of ontologies)
- Updating the demand forecast using latest information
- Responding to peak load or other events that are detected by interacting with consumer.

This entire process implicitly includes a feedback, since any response taken will impact the consumer energy usage, which is measured by subsequent readings of the smart meters.



The technologies that will enable these tasks include scalable stream processing systems, evaluation layer, semantic information integration and data mining systems. The scalable stream processing system is an open-architecture system that manages data from many different collection systems and provides secure, accurate, reliable data to a wide array of utility billing and analysis systems. The system accept meter readings streaming over internet or other communication protocols and detect/react to emergency situations based on defined policies. The evaluation layer captures the raw events and result sets for predictive modeling and sends the information to semantic information integration system. The semantic information integration plays a vital role by using domain knowledge base to integrate and enhance management of transmission and distribution grid capabilities with diverse information and improve operational efficiency across the utility value chain. The data mining systems uses data driven mining algorithms to identify patterns among a large class of information attributes to predict power usage and supply-demand mismatch.

All of these tools will run on scalable platforms that combine public and private Cloud infrastructure, and allow information sharing over Web service APIs while enforcing data privacy rules. A mix of both public and private Clouds is necessary due to data privacy, security and reliability factors. A core set of internal, regulated services may be hosted within the utility's privately hosted Cloud while the public Cloud is used for a different set of public facing services and to off-load applications that exceed the local computational capacity. For more accurate analytics and better demand forecast, the data needs to be integrated with the billing and CRM systems as well. Integrating Billing and CRM systems inside the cloud may prove to be expensive, so in such case it is better to keep the analytics outside the cloud.

Business Value

Smart metering with big data analytics gives an opportunity to focus on accounting and energy auditing, to address theft and billing problems which have vexed the industry. The following lists the Business value that can be realized in using big data analytics:

- Reduce AT&C (Aggregate Technical and Commercial) losses: Enhanced analytics can be used to visualize where energy is being consumed and provide insight into how customers are using energy. It helps to identify the peak load demand and thereby decreasing generation as well as consumption of energy and thus reducing losses. Enhanced analytics also enables the provider to come up with fixed consumption schedule at a fixed price thus reducing commercial losses.
- Analyzing consumer usage and behavior: Big data can be used for enhanced analytics that visualizes where energy is being consumed and provide insight into how customers are using energy. This increases the efficiency of smart grid solutions, allowing utilities to provide smarter and cleaner energy to their customers at an economical rate. A significant amount of value is anticipated to reside in secondary consumer data –behavioral analytics of consumer usage

data will have value to utilities, service providers, and vendors, in addition to the owners (consumers) of that data. Utilities and other energy service providers need this type of consumer data to effectively enlist support for future energy efficiency and demand response campaigns and programs that reward changes in energy consumption. CRM and analytics applications can deliver valuable information to let utilities act as "trusted advisors" to consumers to reduce or shape energy use.

- Manage Load Congestion and shortfall: Analytical tools that processes consumption data can help identify when demand is low or high. With this analysis the utility provider can act upon when to begin shedding load or fire up peaker plants to avoid brown outs and black outs. Long term analysis of grids can provide more detailed information on seasonal and annual changes in both generation and demand, which can be used to model future demand and generation trends.
- In addition to the above benefits smart grid implementation with Big data analytics will play a key role in addressing global issues like energy security and climate change

Data warehousing – Faster and Cost effective

Use case Context

EDW (Enterprise Data Warehouse) are increasingly becoming the lifeline for Enterprise Business. Businesses routinely use the EDW environment to generate reports, gather intelligence about their business and derive strategies for future. Multiple database vendors like Teradata, Oracle, IBM, Microsoft and many others have invested heavily in the field of EDW and have robust products. Till now, these products have been able to serve us well.

All the data in the enterprise, land in the EDW environment in some form or the other. The three characteristics of Big Data viz., Volume, Variety and Velocity poses challenge not just to store the large data-set, but processing and make it available for downstream consumption. As a result the impact of Big Data on the EDW is huge. The current set of EDW products are based on an architecture where it is difficult to scale. The volume of data overwhelms the relational systems which have the concept of a controller (the DB engine). The controller becomes the bottleneck and handling of Big Data becomes expensive. In addition, handling semi-structured and un-structured data are not effectively handled by the current EDW products.

Solution and Architecture

To address the above challenges, we are now seeing a series of innovation in the Big Data space. There are parallel relational data warehouses, shared nothing architecture, DW appliance and MPP (massively parallel processing) architecture. Some key products in the area are Hadoop/Hive, EMC Greenplum, Oracle Exadata, IBM Netezza and Vertica. Each of the solution has infinite scalability and cost-effectiveness at its core.

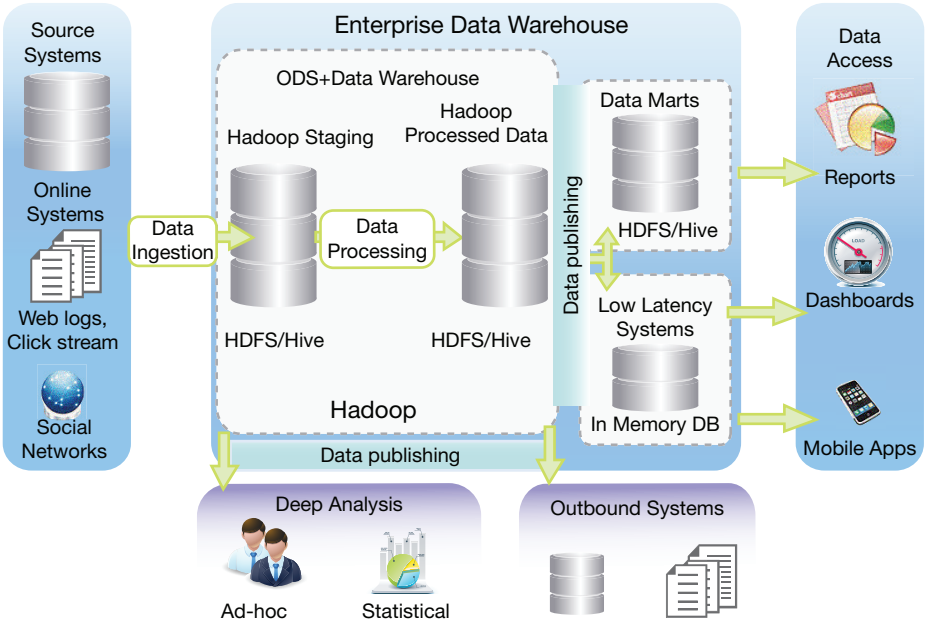
A solution based on Hadoop and HIVE provides a compelling alternative to the traditional EDW environment. Hadoop is a top level Apache Project and was developed by Doug Cutting while working at Yahoo. Similarly, HIVE is another open sourced project and is a data warehouse infrastructure built on top of Hadoop.

Some key benefits of using a combination of Hadoop/HIVE include:

- Cost effective: Both Hadoop and HIVE are open sourced and the initial software cost is zero. In addition, they are designed to run on commodity hardware and the Infrastructure cost is a relatively lesser when compared to conventional EDW hardware.
- Infinite Scalability: Hadoop can scale to thousands of nodes using commodity hardware
- Strong eco-system: Hadoop has now become main-stream and there is a massive support in the industry. We are seeing the eco-system grow at a rapid pace coexisting with landscape.

To address the end to end need of Enterprise Data Warehouse, we need to effectively handle the following:

- Data Ingestion: Data needs to be ingested from a variety of data sources to the Big Data environment (Hadoop + HIVE in this case). The data can either be transactional data stored in RDBMS or could be any other unstructured data that an Enterprise might want to use in their EDW environment



- Data Processing: Once data is ingested to the platform, this needs to be processed to provide business value. Processing can be in terms of aggregation, analytics or semantic analysis. It is interesting to note that, unlike conventional EDW platform, the Hadoop + HIVE environment is well suited to handle the unstructured or semi-structured data. Companies like Facebook, Google and Yahoo routinely process huge volume of unstructured data and derive structured information
- Data Publishing: The processed data needs to be published to a variety of systems for end user consumption. This can include various BI solutions, Dashboard applications or other outbound systems. As support for Hadoop continues to sky-rocket, we are seeing many vendors providing a variety of adapters to connect to Hadoop.

Business Value

The platform provides a compelling alternative to the conventional EDW, especially in the world of Big Data. This kind of architecture is being evaluated by many of our clients. Development of accelerators can help package the above solution as a full-fledged platform and facilitate smoother adoption. Following are some key accelerators that Enterprise can look to develop in the short to medium term:

- Technical Accelerators (Level 0): Big Data Aggregator Framework, Parallel Data ingestion framework, Common Data adapter
- Frameworks (Level 1): Analytics Factory, Semantic Aggregation Framework, Matching Engine, Statistical Analytics framework, Clustering graphs
- Solutions (Level 2): Financial reconciliation, Risk Analysis, Fraud Management, Retail Use Profiling

Q&A

Doug Cutting
Co-founder ,
Apache Hadoop project

Chairman of The Apache Software Foundation and Architect at Cloudera

Doug, first Google, Yahoo and Facebook learned to manage Big Data and now large enterprises have started to leverage Big Data. What are the most common use cases you are seeing in the context of the enterprise?

Most companies are motivated to start using Apache Hadoop by a specific task. They have an important data set that they cannot effectively process with other technologies.

At companies with large websites, this "initial task" is often log analysis. For example, most websites are composed of many web servers, and a given user's requests may be logged on a number of these servers. Hadoop lets companies easily collate the logged requests from all servers to reconstruct each user's sessions. Such "Sessionization" permits a company to see how its users actually move through its website and then optimize that site.

In other sectors, we have observed different initial tasks. Banks have a lot of data about their customers, bill payments, ATM transactions, deposits, etc. For example, banks can combine analysis of this data to better estimate credit worthiness. Improving the accuracy of this estimation directly increases a bank's profitability.

Retailers have a lot of data about sales, inventory and shelf space, that – when they can analyze it over multiple years – can help them optimize purchasing and pricing.

The use cases vary by industry. Once companies have a Hadoop installation they tend to load data from more sources into it and find additional uses. The trends seem clear though: businesses continue to generate more data and Hadoop can help to harness it profitably.

What are the challenges enterprises are facing for the adoption of Big Data?

There's a big learning curve. It requires a different way of thinking about data processing than has been taught and practiced for the past few decades, so business and technical employees need to re-learn what's possible.

IT organizations can also be reluctant to deploy these new technologies. They're often comfortable with the way they've been doing things and may resist requests to support new, unfamiliar systems like Hadoop. Often the initial installation starts

as a proof of concept project - implemented by a business group, and only after its utility to the company has been demonstrated is the IT organization brought in to help support production deployment.

Another challenge is simply that the technology stack is young. Tools and Best practices have not yet been developed for many industry-specific vertical applications. The landscape is rapidly changing direction, but conventional enterprise technology has a multi-decade head start, so we'll be catching up for a while yet. Fortunately, there are lots of applications that don't require much specific business logic; many companies find they can start using Hadoop today and expand it to more applications as the technology continues to mature.

Is Hadoop the only credible technology solution for Big Data management? Are there any alternates? And how does Hadoop fit into enterprise systems?

Hadoop is effectively the kernel of an operating system for Big Data. Nearly all the popular Big Data tools build on Hadoop in one way or another. I don't yet see any credible alternatives. The platform is architected so that if a strong alternative were to appear it should be possible to replace Hadoop.

The stack is predominantly open source and there seems to be a strong preference for this approach. I don't believe that a core component that's not open source would gain much traction in this space, although I expect we'll start to see more proprietary applications on top, especially in vertical areas.

Hadoop started as a sub-project to a search engine and then it became a main project. Now, the Hadoop ecosystem has more than a dozen projects around it! How did this evolution happen in a short span of time?

It's a testament to the utility of the technology and its open source development model. People find Hadoop useful from the start. Then they want to enhance it, building new systems on top. Apache's community-based approach to software development lets users productively collaborate with other companies to build technologies they can all profitably share.

Doug, one last question: Hadoop Creator, Chairman of The Apache Software Foundation and Architect at Cloudera – which role do you enjoy the most?

Hadoop is the product of a community. I contributed the name and parts of the software and am proud of these contributions. The Apache Software Foundation has been a wonderful home for my work over the past decade and I am pleased to be able to help sustain it. I enjoy working with the capable teams at Cloudera, bringing Hadoop to enterprises that would otherwise have taken much longer to adopt it.

In the end, I still get most of my personal satisfaction from writing code, collaborating with developers from around the world to create useful software.

Making it Real – Key Challenges

Context

Digitization of various business functions and adoption of digital channels by the consumers has been resulting in a deluge of information. This is resulting in huge volumes of data getting generated at increasing pace and in various forms and varieties.

Big Data is disrupting value chains for several industries and offering significant business benefits to those organizations which are able to exploit them.

The data volumes are increasing while the costs for storage and processing are reducing and at the same time a whole new set of Big Data technologies like Map-reduce, NoSQL solutions etc have emerged. These technologies are enabling storage and processing of data at higher order of magnitude at much lower costs than what was possible with traditional technologies.

Big Data Challenges

There are several challenges that enterprises are facing in capturing, processing and extracting value from the "Big Data". This section looks into some of the key challenges and the emerging solutions for those challenges. Some of the key challenges include

1. Protecting Privacy
2. Integration of Big Data technologies into enterprise landscape
3. Addressing increasing real-time needs with increasing data volumes and varieties
4. Leveraging Cloud computing for Big Data Storage and Processing

Protecting Privacy

Data mining techniques provides the backbone to harnessing information quickly and efficiently on Big Data. However, this also means there is a potential for extracting personal information by compromising on user privacy (see Sidebar -Privacy Violation Scenarios). In this chapter, we initially describe principles that can be used to protect the privacy of end users at various stages of data life cycle. Subsequently we explore technical aspects of protecting privacy while processing Big Data.

Lifecycle of Data & Privacy

Typically the Big Data lifecycle involves four stages 1) Collection 2) Storage 3) Processing and derive the knowledge and 4) Usage of the knowledge. Privacy concerns can arise in all of these stages. A combination of policy decisions, technical and legal mechanisms are used to address privacy concerns. A brief description of some of the major principles for protecting the privacy of data in its lifecycle is given below.

Privacy Violation Scenarios

- Misusing user password and biometric databases (identity theft)
- Selling transaction databases, credit card databases for monetary gain
- Detecting web access patterns of a particular user from a database of web accesses.
- Identifying a person with a particular disease in a healthcare database.
- Behavioral discovery of a user by correlating activities within a social networking site and also outside
- Setting up monitoring mechanisms to infer behavior patterns of users
- Disclosure of private and confidential information in public
- Command history patterns of a user
- Deep packet inspection of network data to identify personal information like passwords and credit card related transactions
- Exporting sensitive data from a computer through malware, spyware, botnets, Trojan horses, rootkits etc.

1. **Data collection limitation**: This principle limits the unnecessary excessive collection of personal data. Once the purpose for which data is collected is known, collected data should be just sufficient enough for that purpose. This principle is clearly a policy decision on the part of collector.
2. **Usage limitation**: While collecting sensitive or personal data, collector needs to specify what for and how the data is used and limit the usage of collected data for other purposes than the original one.
3. **Security of data**: It is an obligation of data collector to keep the data safe once collected. Adequate security mechanisms should be in place to protect it from breaches.
4. **Retention and destruction**: There is a lifetime associated with the data, once its usage is over the data collected needs to be destroyed safely. It prevents wrong usage and leakage to wrong hands.
5. **Transfer policy**: Often the usage of data is governed by laws which are prevalent at the place of collection and usage. If the data is moved outside the jurisdiction (which is common with the advent of Cloud computing), where the law enforcement is not prevalent in the new place it carries the danger of misuse. Thus either data is not allowed to be transferred or transferred only to those places where similar law enforcement holds good.
6. **Accountability**: When dealing with third party data, the party may ask to designate a person who is the point of contact and take onus of safekeeping, processing and usage of data.

Collection limitation is a policy decision on part of the data collector, usage limitation, securing data, retention and destruction and transfer policy can be addressed by technical means and the last one, accountability is addressed by having a legal team sign a declaration.

Using the collected data for analysis and deriving insight into the data is an important technical step, in the next section we describe some of the privacy preserving data mining techniques.

Privacy Preserving Data Mining Techniques

Objective of Privacy preserving data mining is to reveal interesting patterns without compromising on user privacy. There are variety of techniques used depending on the type of database and type of mining algorithm. Following are the key techniques for privacy preserving data mining.

1. **Anonymization techniques**: Replace sensitive attribute values with some other values. This prevents disclosure of private data. In some cases, this simple replacement alone will not suffice and sophisticated anonymization techniques need to be employed. For example replacing name and address will not suffice, pseudo identifiers like age, sex and social security numbers can be used to identify or else minimize the scope of possibilities. K-Anonymity framework is one good example of this class. In order to reduce the risk of identification, this technique requires that every tuple in the table be indistinguishably related to no fewer than k respondents. There are other techniques like l-diversity model, t-closeness model etc. which fall under this category.
2. **Generalization**: Rare attributes in data items are replaced with generic terms. For example let us consider persons who have a Ph.D degree are very less in employee database. A query which collects user qualifications and their age can be correlated with the result of a query which lists salary and age of the person can reveal the identity of person. This can be prevented by replacing qualification (Ph.D) with another generic term called graduation which makes it difficult to correlate and infer.
3. **Randomization**: In this technique noise is added to the fields of records. This prevents retrieving correct personal information however the aggregate results are preserved. For example salary and age of employees is randomized however queries like average age and average salary will result in correct reply. One of the advantage of randomization techniques is, it can be used with individual records (noise can be added at the time of collection of data) and does not require knowledge of other record values, hence this method is more suitable for data which is generated as a stream. There are variety of randomization techniques such as additive randomization, multiplicative randomization and data swapping techniques in this class.
4. **Probabilistic or no results for queries**: Modify the query results which potentially compromise user privacy in such a way that, rather than giving exact results they either give probabilistic answers or null results.

Privacy Preserving Data Publishing

Once the data is processed and insight is gained about the data, in some cases the knowledge is shared either publicly or to limited audience. Precautions should also be taken to prevent misuse of such knowledge. One of the key questions while sharing results is - Is it the raw data or just the inference? If raw data is shared then adequate care should be taken to mask the data so that analysis is reproducible however individual identities are hidden. To a major extent publishing data is bounded by legal means. For a snapshot of privacy laws look at the Sidebar-Data Privacy Protection Laws

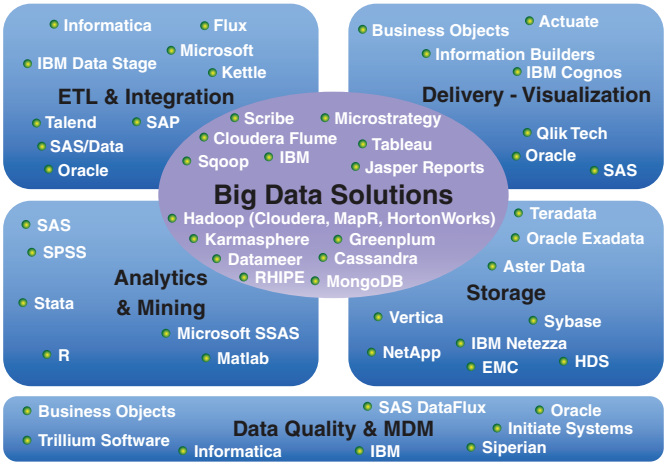
Data Privacy Protection Laws

- Payment Card Industry Data Security Standard (PCI DSS): Defined to protect financial transaction data from potential breaches
- Health Insurance Portability and Accountability Act (HIPAA): This US law regulates the use and disclosure of Protected Health Information held by "covered entities"
- Sarbanes–Oxley Act (SOX): Is a standard for public boards and public accounting firms in US
- Personal Data Privacy and Security Act: Deals with prevention and mitigation of identity thefts.
- US privacy act: Limits the collection of personal data from US federal agencies.
- Data Protection Directive (95/46/EC): An European law which mandates the processing of European citizens personal data within Europe.
- UK data protection act: Serves the same purpose US privacy law does.

Integration of Big Data technologies into enterprise landscape

Enterprise Data Warehousing (EDW) and Business Intelligence solutions form an integral part of business decision making in enterprises today. Large enterprises typically would have one or more of these products already in use. Emerging Big Data technologies and solutions are largely complementary to some of these and sometimes provide alternatives addressing extreme volumes or velocity or variety requirements at lower price points. A key challenge is determining where Big Data technologies fit in a typical enterprise and how they are used in conjunction with all other existing products in the enterprise. A graphical representative of some of the typical BI/EDW capabilities and representative players is shown below. (Note: It is a representative list, not comprehensive and not indicative of any rankings and also there are vendors/solutions that cut across several functions)

The figure gives an idea of where the Big Data Solutions fit in the current enterprise context. Big Data technologies are today being used primarily for storing and performing analytics on large amounts of data. Solutions like Hadoop and its associated frameworks like Pig, Hive etc. help distribute processing across a cluster of commodity hardware to perform analytic functionalities on data. Hadoop based data stores as well as NoSQL data stores provide a low cost and highly scalable infrastructure for storing large amounts of data.



Challenges with Leveraging Big Data Technologies in Enterprise Landscape

Enterprises that want to adopt Big Data solutions have been facing a number of challenges in getting these tools to integrate with the existing enterprise BI/EDW/ Storage solutions from vendors like as Teradata, Oracle, Informatica, Business Objects, SAS, etc. Some of the challenges include:

Data Capture and Integration: Lack of a proper ETL tools that can load data from existing data sources into a big data solutions like Hadoop Distributed File System, Cassandra, MongoDB, GraphDB etc.

Enterprises have large amounts of data stored in traditional data stores including file system, RDBMS, DW systems, etc. To use this information to derive any useful analytics out of them, it is required to load this data into the Big data solutions. Most of the ETL systems do not support such bulk loading of data from traditional data sources to the options available in Big Data space. There are several specialized open source solutions like Key-Value stores (Cassandra, Redis, Riak, CouchBase) , Document Stores (CouchDB, MongoDB, RavenDB), Big Table/Column Stores (HBase, HyperTable), Graph databases (Neo4j, GraphDB) etc but the solutions for integration into enterprise data stores like CRM, ERP systems etc. are very limited.

Lack of Data Quality: Traditional Data Quality solutions provided by vendors like IBM, DataFlux, Business Objects provide extensive capabilities for metadata

management and addressing data quality issues. However there is limited integration of these with the Big Data technologies. This is resulting in lot of custom solutions for scenarios where big data technologies are used.

Richness of Analytics/Mining capabilities: Big Data solutions and frameworks available today for analytics like Apache Mahout provide a limited number of algorithm implementations and their usability is also limited compared to the kind of features business analysts have been used to with commercial solutions.

Limited Data Visualization and Delivery capabilities: There is limited support for visualization of analysis results in existing Big Data solutions. A major requirement for business users is ability to view the analyzed data in a visually comprehensible manner. The BI/DW reporting solutions allow users to generate these visual charts and reports by connecting with traditional BI solutions easily. Support for Big data solutions such as Hive, HBase, MongoDB, etc. in such popular reporting tools in limited at this point of time.

Limited integration with Stream/Event Processing solutions: Several Big Data frameworks like Hadoop provide good results for batch requirements but they are not architected for real-time processing requirements. There are several solutions like CEP which address real-time processing needs but their integration with Big Data solutions is limited.

Limited integration with EDW/BI products: Traditional BI/EDW solutions provide advanced features like OLAP enabling easy slicing & dicing of information and also enabling users to define and analyze the data through a user friendly UI. This allows business analysts with limited technical expertise use these solutions to address business requirements. The user experience maturity aspects are still in very early stages with the Big Data Solutions available currently. A lot of work has to go in making them more user-friendly.

Emerging Solutions and Where to Start

Most of the initial work in developing 'Big Data' technologies and solutions that help manage extreme data volumes were from internet giants like Google, Yahoo, Amazon, Facebook etc. who later open sourced those solutions and now there are companies like Cloudera, MapR, Hortonworks, DataStax, etc. providing commercial support for them. Driven by increasing adoption of these solutions, a number of established enterprise players with offerings in the BI/EDW/Storage space have now started offering/integrating Big Data solutions with their product stacks.

In this section, we look at some emerging solutions to overcome the challenges in integrating big data solutions with existing enterprise solutions.

Big Data Integration support from ETL Tools: Several ETL tool vendors like Informatica, Microstrategy etc. have started supporting 'Big Data' solutions like Hadoop/Hive etc.

A common requirement in integration is for extracting data from online RDBMS data stores into frameworks like Hadoop for further processing. A commonly used pattern for these situations is to first use the export capabilities provided by data store solutions or using ETL tools like Informatica to extract the data into flat files. Then in next stage, frameworks like PIG are used to then load the data into Hadoop Distributed File System (HDFS) or into NoSQL data stores. After that, frameworks like Map Reduce are used for processing and aggregation operations and finally the result of the processing is loaded into a RDBMS or a NoSQL data store.

There are also emerging frameworks like Flume, Scribe, Chukwa that are designed to collect data reliably from multiple sources, aggregate and load them into the 'Big Data' stores.

Big Data Integration support with Visualization Tools: Big Data solutions are also used for several scenarios like analysis of sales, product and customer transaction information from existing data sources such as RDBMS, flat files to generate aggregation and projection reports. A challenge here is to integrate with an existing Visualization tool to show the projection output through user-friendly charts.

Reporting solutions such as Tableau, JasperReports and Pentaho are providing support for directly connecting with different big data stores and generating charts/ reports.

There are still some challenges like responses from a Big Data store may be too slow for interactive analysis or the features available maybe limited. A commonly used solution pattern to get around this challenge is to use frameworks like Hadoop for distributed processing, analytics and store the results in a traditional RDBMS databases so that existing visualization tools can be used.

Big Data Integration support for Analytics Tools: Big Data solutions enable processing and analytics at large scale in several scenarios. For example - Large ecommerce web sites having millions of customers are expected to provide a personalized surfing experience to each visitor. Due to the large amount of data involved, Big Data solutions are used to retrieve user transaction details in the web logs and identify user preferences. Based on that, items are recommended to the user in near-real-time. A challenge here is to integrate commercial Analytics products with a Hadoop or NoSQL based data store to perform the necessary analytics.

A number of commercial BI tools such as Teradata Aster MapReduce platform, IBM BigInsights are integrating Big Data processing frameworks like MapReduce into their products. There are some initiatives underway in open source solutions space like RHIPE that are looking to integrate known analytics packages like open source R with Big Data solutions like Hadoop but these are still in early stages.

Addressing increasing real-time needs with increasing data volumes and varieties

Another key challenge is addressing the increasing needs for real-time insights and the need for interventions based on those insights at real-time. For controlling process efficiencies, business effectiveness, it is becoming more and more imperative to make decisions "on the fly" based on, "on the fly" data. Here are some of the scenarios with such needs:

Online Commerce - Analyzing online customer behavior in real-time and providing personalized recommendations based on identified customer preferences.

Location-based Services — Location information of millions of mobile subscribers needs to be tracked continuously and this needs to be combined with customer profile and preference information in real time to generate personalized offers related to the local businesses.

Fraud detection – Financial Services firms are developing applications that detect patterns indicative of fraud based on analysis of previous transaction history and check for these patterns in real-time transaction data to prevent fraud in real-time.

Network Monitoring and Protection —Prevention of malicious attacks needs continuous monitoring of application and network data in real-time and reacting to suspicious activity.

Market Data solutions- Financial Services firms need to analyze external market data in real-time and arrive at recommended financial transactions before the opportunity disappears.

Social Media–Twitter, Facebook or Smartphones based messaging can help reach out to trillions of subscribers in a second to market products or services. Analyzing these to identify key influencers and targeting them requires processing large volumes of data in real-time.

All the above scenarios throw significant challenges from Volume, Variety and Velocity of data perspective.

Challenges with real-time needs

Some of the challenges and considerations are,

Capture and Storage: The sources for the information are disparate such as devices, sensors, social channels, live feeds. Most of the times, the data is in the form of message streaming from these sources. The volume of information to be stored and analyzed could be huge. The variety of messages could also be very different where one message may not be in context or related to next arriving message. The velocity of messages coming in could be as high as million messages per second.

Processing and Analytics: The processing of data needs to happen in real time. Processing and establishing patterns amongst messages involves complex computations such as detecting and establish patterns among events (correlation), applying rules, filter, union, join, and trigger actions based on or absence of events, etc.

Result Delivery: After processing, the information needs to be presented to the end user in real time in the form of appealing Dashboards, KPIs, Charts, reports, email and the intervention actions like sending alerts using user preferred channels such as Smartphone, Tablets, or Desktops (Web, thick client).

Reliability, Scalability: Systems that process such information need to be highly fault tolerant where loss of data due to missing out on one message may become unaffordable at times. They also need to be scalable and elastic so that they can scale easily to cater the increased demand on processing.

Due to all the above challenges, performing analytics and data mining on Big Data in real- time differs significantly from traditional BI because in real-time it is not feasible to process the messages and derive insights using the conventional architecture of storing data and processing in batch mode.

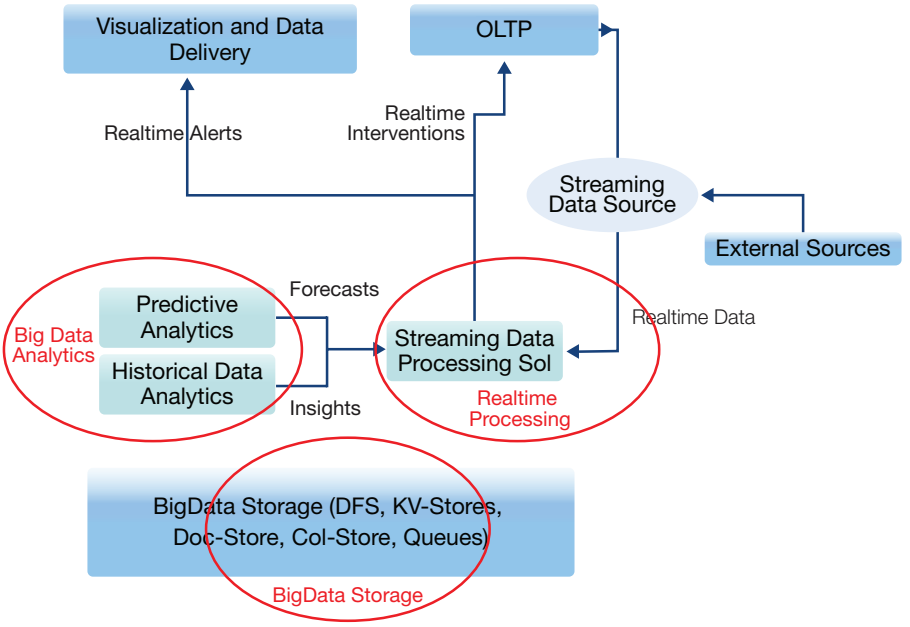
Emerging Solutions and Where to Start

To address above challenges, enterprises need to look at real time data processing solutions too as part of their enterprise Big Data strategy.

Store and Analyze Solutions: Data is mined and analyzed for historical patterns using a combination of emerging Big Data technologies like Hadoop using MapReduce architecture pattern and traditional BI solutions. Future trends and forecasts are determined using predictive analytics techniques. All this is performed on data collected over a long period of time. Solutions like Hadoop help store and analyze large volumes of data but they are not designed for real-time response needs.

Stream/Event Processing Solutions: Processing streams of data with real-time response needs can neither be handled through Historical Analytical DW nor Hadoop based architecture because of the challenges mentioned earlier. Hence the processing of such streams is done using stream centric solutions like Complex Event Processing (CEP) solutions. Complex Event Processing (CEP) is continuous and incremental processing of event streams from multiple sources based on declarative query and pattern specifications with near-zero latency.

By combining these two techniques, namely Store and Analyze and Stream processing, the requirements for processing and analyzing large amounts of data over a long period of time while at the same time generating insights, forecasts and acting on them in real-time can be achieved. Historical patterns and forecast from the first stage provide inputs for the second stage which can help apply them in real-time.



Conceptual solution to address Velocity and Volume requirements simultaneously

Stream/Event Processing Solution Architecture

A typical Real time Streaming Solution has three key components.

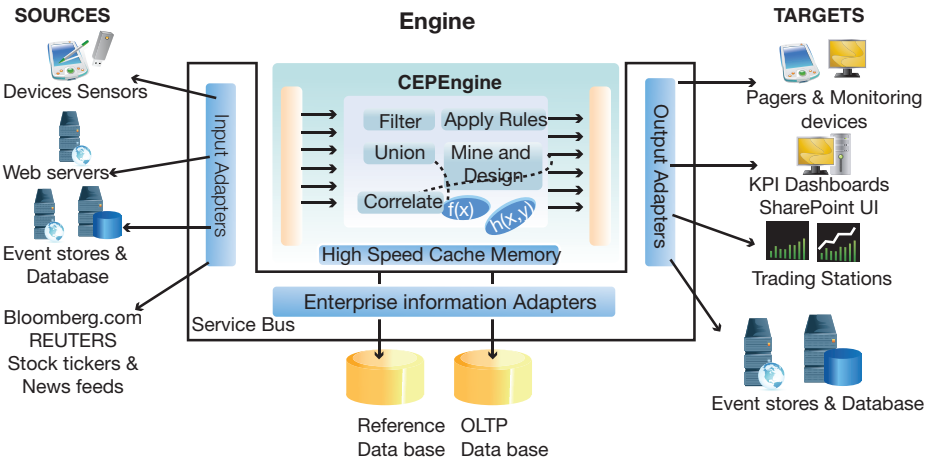
Data Sources –Messages coming from diverse data sources such as devices, sensors, social channels, live feeds needs to be captured. Most of the times, the message coming from same source do not have standard and consistent schema or structure. Input adapters help capture and transfer these messages from these source systems.

Engine – The CEP engine needs to process the stream of messages in real-time by applying mathematical algorithms, functions, filters, business rules, establishing patterns at the same time combining the insights or reference data from enterprise information data stores. The engines generally use parallel processing, multi-threading, very high speed large memory cache, and highly optimized algorithms to perform all of this in the flight.

Targets- Targets owns the responsibility to present the processed information through different delivery channels. Output adapters are used to connect to these varied presentation channels. Individual presentation channel owns the responsibility of rendering the information in the required form factor in visually appealing manner.

StreamBase, Tibco ActiveSpaces, IBM WebSphere Business Events, Sybase Aleri, ruleCore, ActiveInsight, Microsoft StreamInsight are some examples of commercial CEP solutions. BackType-Storm, S4, Esper are examples of open source solutions.

As discussed, significant number of business use cases cannot be addressed through the conventional Enterprise Big Data BI architecture strategy because of the unique challenges it offers around various velocity characteristics. To deal with this, Complex Event Processing (CEP) architecture is useful. However, CEP cannot be looked in isolation from Enterprise Big Data BI architecture and needs to be seamlessly integrated in such a way that Complex Event Processing is performed in the context of enterprise Big Data.



Real time CEP platform Architecture

Leveraging Cloud computing for Big Data Storage and Processing

Cloud computing and Big Data are two key emerging disruptive technologies; these are accelerating business innovation and enabling new, disruptive solutions. The adoption footprint of these two most disruptive entrants has been a global phenomenon those are cutting across multiple industry verticals, and geographies and rate of their adoption has been quite fast in today's enterprise landscape.

Enterprises are seeking answers to some of their key business imperatives through Big Data analysis like – Modeling true risk, customer churn analysis, flexible supply chains, loyalty pricing, recommendation engines, ad targeting, precision targeting, Point-Of-Sale transaction analysis, threat analysis, trade surveillance, search quality optimization, and various different blended mashups such as location, context, time, seasonal, behavioral ad targeting. To address these business requirements a "Data Cloud" with an elastic and adaptive infrastructure such as public and private cloud platform for enterprise data warehousing and business intelligence functions is being considered. Cloud computing and Big Data together allow analysts and decision makers to discover new insights for Intelligence Analysis, as demonstrated by Google, Yahoo, and Amazon. By leveraging such "Data Cloud" platforms to store, search, mine, and distribute

massive amounts of data, businesses are now enabled to get answers to their ad-hoc analysis questions faster and with more precision.

The following table outlines a few important types of analytics those are performed on Big Data and in many cases, can effectively leverage "Data Cloud" and "Cloud Infrastructure / Utility Cloud":

Types of Analytics	Characteristics	Examples
Operational Analytics	• Complex analytic queries • Performed on the fly as part of operational business processes, • Concurrent, high data volume of operational transactions	Real-time fraud detection, Ad-serving, High Frequency Trading
Deep Analytics	• Typically multi-source, • Non- operational transactions data • Complex data mining and predictive analytics, • Real-time or near-real-time responses • Uses Map Reduce type framework, columnar databases and in-memory analysis	Gaining insight from collected smart utility meter data
Time Series Analytics	• Analytics with the concept of a transaction: an element that has a time, at least one numerical value, and metadata	Algorithmic Trading
Streams Analytics	• Ultra low latency analysis of data • Applying transaction-level logic to real-time observations	Logs Analysis, Web crawling, Recommendation Engine
Insight Intelligence Analytics	Analysis over a vast complex and diverse set of structured and un-structured information	

A large set of data now exists in the cloud (private, public and hybrid) space and many organizations and applications are also making their way into cloud platforms every day. So, some key questions that enterprise architects are facing is how to leverage the cloud computing infrastructure for Big Data storage and processing. Is it possible to use public cloud platforms for Big Data analytics? What are the benefits and what are the challenges?

Big Data in Cloud: Benefits and Challenges

The table provides an overview of the different architectural options for "Data Cloud", the considerations and the pros & cons:

	Options	Implications	Pros & Cons
Public Cloud	Leverage Public Data PaaS Frameworks like Amazon Elastic MapReduce	• Easy to get up and running. Ex: Amazon map reduce jobs (EMR, EC2) • External data upload requirement poses security concerns • Significant size and Processing times impacts latency	+ Easy and Low cost + Can be used to study the map reducibility of the problem – Limited Public Data PaaS options available – Vendor lock-in – Perimeter security concerns, data transfer speed (more time to run) and high latency limitations apply
	Leverage Public IaaS and setup Big Data Solutions on them	• Requires on-premise Map Reduce Cluster • Data transfer requirement remains same	+ No procurement lead times + No vendor lock-in + May be cloud-neutral – Perimeter security concerns, data transfer speed (more time to run) and high latency limitations apply
Private Cloud	Build the Map Reduce clusters on a set of virtual machines	• Provision VMs on shared infrastructure • Setup the Map Reduce clusters on the shared storage infrastructure • Overheads of virtualization and limitations of architecture	+ Easier Provisioning + Dynamic scale out of infrastructure – Shared infrastructure and virtualization over heads may impact on latency
	Build out the Map Reduce clusters on its own dedicated hardware	• Provision the dedicated Hardware • Build out Map Reduce clusters on a share nothing setup	+ High performance, low latency – Increased Cost with lower utilizations of dedicated hardware – Procurement lead time

The key benefits from leveraging cloud infrastructure for Big Data needs are

Low costs: Using Infrastructure via utility cloud model there by reducing infrastructure costs

Fast Turn- around Time for Infrastructure: On-demand provisioning of cloud infrastructure reduces the lead times

High Elasticity and scale-out: Massive Parallel Computing using Data Cloud with dynamically scalable infrastructure

The key challenges are:

Data Transfer Limitations: Transferring data is key challenge especially with public clouds. Most enterprises still have majority of their system on-premise so the data generated is mostly in those systems. Latencies in data transfer and costs will be key limitations.

Security and Privacy Challenges: Moving data to a public cloud infrastructure will result in associated security, privacy challenges as the data is moved out of enterprise boundaries.

Performance Challenges: Virtualization overheads and limitations of deployment options with cloud infrastructure can lead to performance degradation resulting in need for more infrastructure footprint.

Emerging Solutions and Where to Start

Most of the challenges in leveraging cloud infrastructure for "Big Data" solutions are more fundamentally related to cloud adoption. There are several solutions being developed to address these cloud challenges like in cloud security area, solutions like Virtual Private Clouds, Federated Cloud Identity and Access Management solutions etc. As cloud adoption increases and as more and more enterprise workloads move to cloud, it will make more sense to perform Big Data processing on cloud.

Public Datasets: Currently, there are several scenarios where data is publicly available like the huge datasets being put out by the governments as part of their transparency and citizen involvement initiatives. There are also various other sources of public Big Data sets like market data that can be processed in financial services, genome data from projects like 1000 genome project which can be analyzed for drug discovery, weather data that can leveraged in agriculture and transportation domains etc. Cloud based Big Data platforms are ideal for such scenarios where security concerns are a non-issue. Several public cloud providers are also hosting such public datasets like for example the AWS Public datasets on AWS which reduces the complexities involved in transferring large amounts of data and also reduces the costs.

External Collaboration: Innovation is moving towards ecosystem driven models where multiple partners collaborate to co-create. Such B2B exchanges also offer opportunities for "Big Data" on public clouds.

Q&A

S. Gopalakrishnan (Kris)
Co-Chairman, Infosys Limited

Kris is on the Global Thinkers 50 and Co-Convener, ICT & Innovation, World Economic Forum

Is "Big Data" a hype or is it a transformational change?

Any disruptive change comes with a period of hype. In the past, IT was limited to Business back office automation, scientific research computation, personal productivity. With the advent of internet web, IT connected Businesses, and now it is connecting People.

In the future, software is going to be everywhere - Mobile devices, Medical equipment, Electric grids, Home Appliances, Vehicles etc. This is already happening and will continue to grow in scale.

The outcome of this is an explosive growth of data with ubiquitous compute, storage and bandwidth. It is becoming more important to make sense out of this data and gain competitive advantage. This trend is reflected across Industries.

Big data technologies along with Cloud, Mobile and Social will help companies to differentiate through productivity and efficiency improvement.

Your view on how large enterprises can harness this change?

I will give you few examples.

A financial service provider, could do a credit and operational risk analysis across product lines. Also, they can perform large scale transaction analysis for fraud detection and risk management.

Internet companies like Google, Amazon, Yahoo pioneered this Big Data management in their search, personalization platforms. Large enterprises are also piling huge amount of data from which they can derive insights and make informed decisions.

Questions like, what consumers are saying about a brand/product? What consumers want? Such needs can be analyzed in near real time using Big data distributed computing platform.

How do you see the connection between Big Data and Infosys strategic theme of Building Tomorrow's enterprise?

Big Data management is a great enabler for Building Tomorrow's Enterprise.

Digital consumers now are active, informed and assertive. Big data analytics helps in delivering greater personalized products and services to them.

Sustainable tomorrow needs efficient management of Power, higher yield in Agriculture etc. These use cases require large and complex datasets processing.

An outcome of Pervasive Computing is the ability to gather large amounts of information from sources like social networks, mobile location based services, online shopping. This information with Big Data analytics helps enterprises to become more intelligent.

In the health care, Drug development analysis, genome analysis, bio-informatics all requires Big Data management.

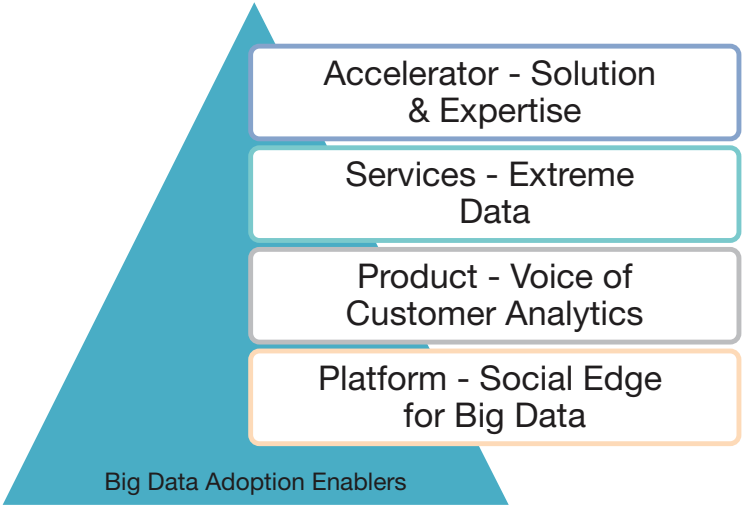
Thus Big Data will become one of the key enabling technology for Building Tomorrow's Enterprise.

Making it Real – Infosys Adoption Enablers

Big Data solution development is a challenge because of considerations like Extreme data volumes, wide variety of data sources, formats and complex integration mechanisms and the growing need for real time actionable insights and meaningful information from all this, while the technologies are new and still evolving.

Significant engineering efforts in terms of partitioning and sharding are required to scale these solutions beyond the specified range. Large scalable MPP solution tend to fail when we are doing joins across tables that span about 100 TB. A new set of Extreme best practices, reference architectures, processes, frameworks, products and platforms are needed to address the Big Data challenges. The elastic compute capability required by these solutions makes them an ideal solution to be implemented in the cloud.

Following are the key enablers offered by Infosys in Big data Adoption:



The real value from Big Data solutions for the enterprise is the actionable intelligence gained by analyzing large volumes of data within and outside the enterprise. The objective of this is to alleviate the concerns around data ingestion and data consumption allowing the customer to focus on the value added activity of data processing.

Accelerators – Solution

Infosys Big Data Solution Accelerators : A Big Data Solution Accelerator Framework that captures and provides reusable artifacts like process artifacts, design artifacts, code components and horizontal and vertical solution frameworks for the various

Big Data solution development life cycle stages shall help reduce the risk of failure for Big Data project execution. The vision for the Infosys Extreme Data Accelerator Framework (InDAeX) being developed at Infosys Labs is described below:

Inception Stage Accelerators

The platform provides accelerators for the initial stages of Big Data projects like Big Data ROI estimation tool, Cost and Time Estimation templates with common tasks.

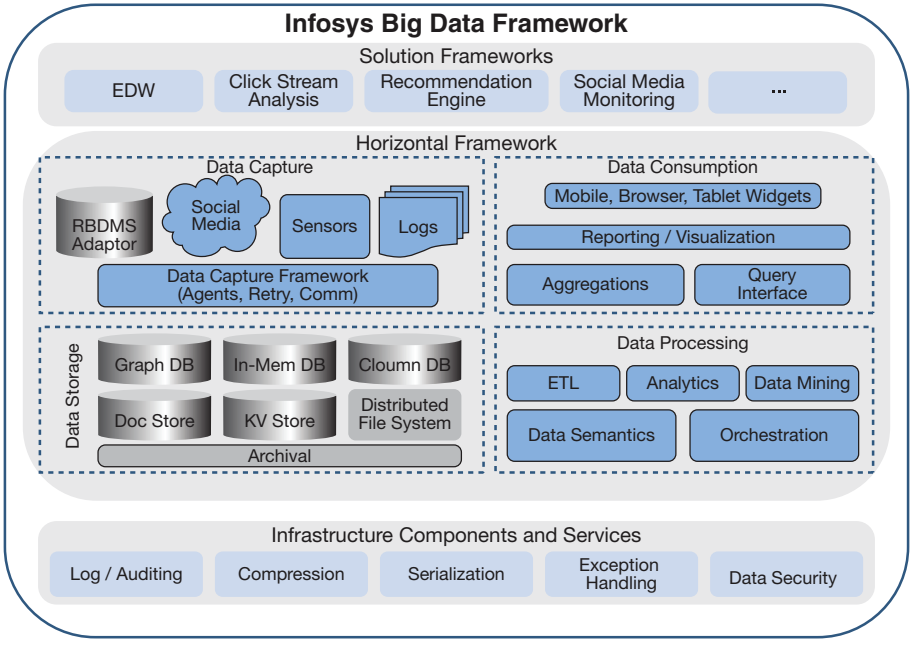
Architecture & Design Stage Accelerators

The platform consists of Reference architectures for common Big Data solution patterns like Peta-byte scale Enterprise Data warehouse, Real time BI, Log Analysis, Social Media Monitoring, eCommerce Recommendation Engine etc.

Build Stage Accelerators

The platform consists of execution methodologies and best practices like

- Big Data Migration methodology with data security and privacy, reliability of data ingestion and access in the face of unreliable networks to the data sources.
- Performance Engineering Methodology
- Trusted App Development Methodology
- It also consists of building blocks and tools like



- Reusable Infrastructure components like Data compression, Data storage helpers, Data Integration adapters, Data processing utilities like object converters, scripting components, Data visualization components.
- Horizontal frameworks that address data capture, processing, storage and data delivery concerns
- Vertical frameworks for the various solution patterns
- Tools like code generation tools, testing tools.

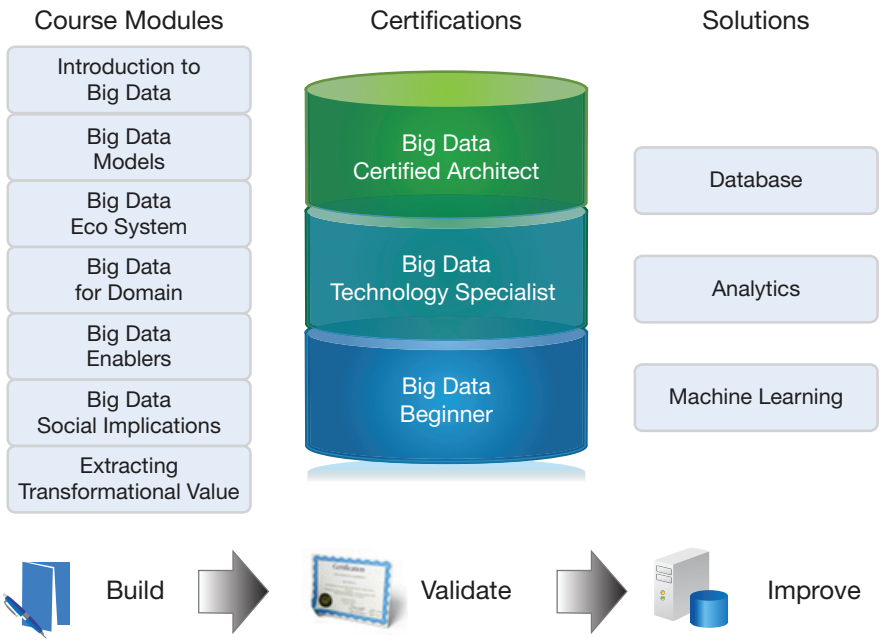
Operations Stage Accelerators

The platform consists of accelerators for post-production operations like management and monitoring tools.

Accelerators – Expertise

At Infosys, Certification is the professional goal and one of the ways to test one's knowledge and measure oneself w.r.t global benchmark. Infosys Big Data Certification Program (IBCP) was launched to validate the technical, application and solution skills of the learnings that were derived from the project implementations done in the area of Big Data.

IBCP has three levels namely- Big Data Beginner Certification, Big Data Technology Specialist and Big Data Architect



Benefits:

- Build on the knowledge of Big Data to stay current on the technology.
- Validate your knowledge on Big Data to know where you stand in technology.
- Improve Solution Capabilities and become Trusted Advisors to our Clients.

Services – Extreme Data

Extreme Data services are focused on providing a complete solution to the customer's data processing problems using a combination of Big Data and traditional technologies. With the evolution of Big data, caused by unprecedented and unpredictable growth, which is in the magnitude of 'Peta Bytes' needs to be handled. This enormous amount of data gets complex due to 'Unstructured-ness' and Frequent changes in data model, compounded by necessity to capture adhoc Information.

A key aspect of the Big Data solutions for a customer needs to take into consideration that the Big Data solutions are not an end in itself but needs to fit into the data supply chain and data processing capabilities that exist in the company. It is essential to take an end to end solution perspective leveraging multiple big data and traditional solutions to solve the customer problem.

It is essential to understand that the big data solutions are not a replacement but a compliment to existing data solutions to help solve the technological problems related to the three V's of data i.e.

1. Volume: Processing very large volumes of data
2. Velocity: Processing large amounts of data flowing in from sensors and other ubiquitous computing solutions
3. Variety: Processing various complex data types e.g. audio, video, seismic, geo spatial etc.

The types of big data problems the customer needs to solve can be classified into two buckets

1. Solution to a functional or business problem e.g. making better recommendation to customers on their website to improve the sales completion percentage by 20%
2. Technical solutions that leverage big data include improved ETL processing, unstructured data processing and cost effective secondary storage for data.

There are two distinct patterns to the Big Data solutions that customers implement today

1. **Large volume data processing**: This is where large volumes of data both internal and external to the enterprise have been accumulated and we need to mine this data for actionable intelligence.
2. **High concurrency, low latency access to data**: This is typically the case where a large number of customers are trying to read and write onto a dataset.

Implementing the Big Data raises three fundamental questions around dealing with data at such a large scale:

1. How do you process this data? Given the volumes traditional relational solutions fail and how do you leverage the newer solutions like Hadoop to provide solutions that fit in with the data strategy of the enterprise?
2. How do you manage this data? Data be it big or small is a corporate asset and how do you manage this data especially when the scale is large? What is the lifecycle of big data?
3. How do you integrate, migrate and synchronize big data?

Three Extreme Data services are designed to address various aspects:

1. Extreme Data Processing:
    a. Implementation of Extreme data processing platform
    b. Using Hadoop, Exadata, MongoDB, Cassandra etc.
    c. Low Latency, high concurrency data processing
    d. High Volume data processing
    e. Big Data processing in the cloud
2. Extreme Data Management
    a. Enterprise Data Marts
    b. Data virtualization & federation
    c. Master Data Management
    d. Metadata Management
    e. Data Archiving to the cloud
    f. Extreme Data Migration
    g. Database Consolidation under virtualized environment
    h. Large Scale Data movement between RDBMS and BIGDATA
    i. Data Integration and synchronization across solutions
    j. Infosys Services around 'Extreme Data'
3. Extreme Data Advisory Services (Services geared towards helping client identifying its Extreme Data need and the solution for the same)
    a. Helping client identifying the real 'Extreme Data' need in terms of volume and usage
    b. Architecture definition integrating traditional RDBMS/DW solution and 'Big Data' solution
    c. Architecture definition for 'Extreme Data' solution based DW

d. Technology Selection through Qualitative and Quantitative analysis

e. Roadmap creation - Defining Enterprise Adoption plan

4. Extreme Data Development Services (Services geared towards Implementation of identified the Extreme Data solution)

   a. PoC service to identify the integration challenges and prove the feasibility of regular feature requirements

   b. Prototype creation helping people understand the 'Extreme Data' solution and challenges

   c. Key Technical Component Development abstracting complexities of 'Extreme Data' technology from regular developers

   d. Performance Engineering

   e. Infrastructure Sizing

   f. Development and Deployment automation

   g. Program Management

5. Extreme Data Migration Services (Services geared towards helping client in moving from current solution to a desired one)

   a. Defining Migration Path from current to future state

   b. Assessment of current solution w.r.t 'Extreme Data' need

   c. Automation for data migration ensuring data quality and sanity

   d. Infrastructure Sizing for target platform

   e. Program management

6. Extreme Data Sustenance Services (Services geared towards ensuring optimal TCO in maintaining 'Extreme Data' solution)

   a. Priming 'Extreme Data' solution for Production support, Professional support and Solution enhancement insulating client from the product vendor

   b. Maintenance service validating product enhancements and fixes and Ensuring optimal use of product features and capabilities.

Client success stories of Infosys' Extreme data services:

Following are some of the case studies showcasing the capabilities which Infosys has in Extreme data services.

Case study 1 : The client is one of the largest manufacturers in the US. They have a problem with exponentially growing data with the data volume expected to reach 10 Peta bytes by 2014. The cost of infrastructure to support this growth using the traditional solutions is also growing exponentially.

Infosys defined and implemented a Big Data solution combining traditional data warehousing components i.e. Teradata with Hadoop to create a solution that is cost effective and leverages the existing investment in the BI solutions.

Solution:

- Hadoop as the data warehouse platform (10PB data capacity)
- Informatica extracts the data from source systems and loads into HDFS
- ETL and Analytics done on Hadoop cluster (Hive, Pig, MapReduce)
- Results of the processing is sent to Teradata for reporting and creation of cubes for analysis
- Reporting needs (BI tools like Business Object, Corda) supported by Teradata
- 330 Nodes Hadoop cluster
- Ability to load 1TB/hour

Benefits

- Significant cost savings
- MAD(Magnetic, Agile and Deep dive) approach to solution design
- Leverage existing investment in Teradata infrastructure.
- Ability to handle exponential data growth
- Ability to process structured as well as unstructured data under a single platform
- Effective handling of dynamic schema and frequent schema changes

This implementation is the template for an efficient and cost effective Data Warehouse solution using Big Data solutions. This also indicates that the future of Information Management is going to have Big Data as one of its pillars and this new world will be built using MAD architectures.

Case Study 2: The client is one of the largest retailers in the US. For their business unit they needed to build the ability to crawl competitor websites and adjust their product pricing based on the crawl information.

Infosys implemented a solution using Hadoop and the grid computing solution Condor. The combined solution was able to do the same crawls about 60 times faster. The business benefit from this was the ability for the company to crawl additional retailers' sites thus helping to converge on the most competitive price for their product.

Solution:

- Condor (Grid computing framework) to manage Crawl farm (over 200 machines)
- Hadoop to process the output from the crawl
- Azkaban to manage the workflow

- Parse and extract useful data from crawl data using Hadoop MapReduce solution using Pig
- Store extracted data into a distributed database
- 60 TB Hadoop Cluster.
- Hive for reporting and ad-hoc analytics
- Average size of crawl between 50MB to 1.5TB based on the type of crawl.

Benefits:

- Complete evolution in terms of crawl performance capability
- End to end visibility of crawling and processing of data
- Mahout based approach to Item information matching
- Significantly increased the number of crawls significantly impacting the confidence in pricing strategy

Case Study 3: Very large scale (over a billion transactions) reconciliation engine for a large US Manufacturer using Mongo Db

Solution:

- Mongo Db used to match millions of entities in less than an hour using commodity hardware based horizontally scalable architecture
- GridGain used for distributed computation at the Java layer
- Using principles of platform as a service model to onboard tenants through configuration
- Use of Mongo Db's secondary indexing capabilities to do exception management in a fast and effective way

Benefits:

- Running Matching programs in parallel for Millions of records to finish the runs in stipulated time
- Support for Horizontal Scalability
- Availability and Durability
- Secondary Index to search across multiple data sources and fields
- Lower TCO
- Ability to add/remove nodes based on seasonal workload

The knowledge gained through these projects along with the broader patterns of Big Data processing developed at Infosys is leveraged in delivering Extreme Data services.

Product – Voice of Customer Analytics

Understanding customer behavior and improving customer experience is more challenging than ever in today's digital economy. Today's digital consumers connect to the enterprise through multiple touch points and also express their views/ opinions in Social Media sites like Twitter and Facebook. While there is no debate that data generated at these touch points contain all the ingredients to manage and improve customer experience – challenge always has been in the limitations of existing technologies to exploit the data in a scalable and cost effective manner. With enterprises beginning to adopt Big Data Technologies – it is now feasible to process the huge volume of data from the touch-points and social media and co-relate with traditional transaction and demographic data to derive deeper insights into customer interests and pain-points.

Need for Big Data Adoption Enablers

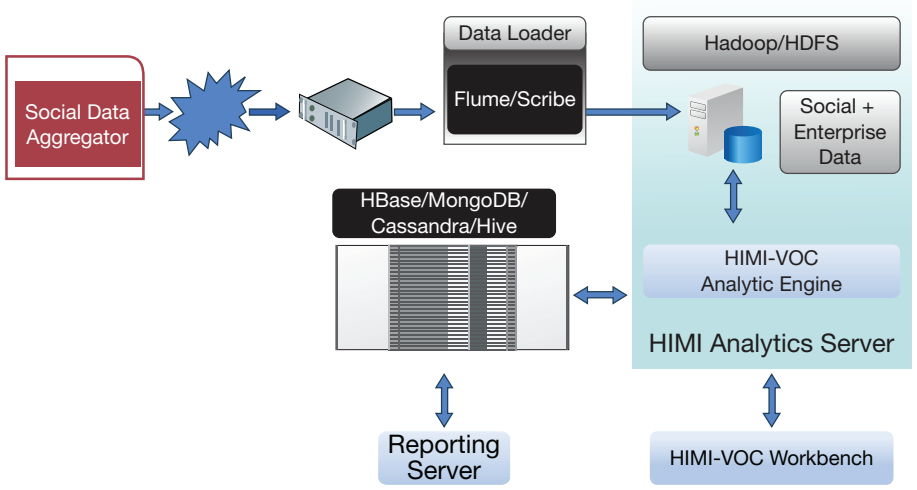
As Enterprises start IT programs to build out the infrastructure and skill sets and applications on Big Data Technologies, there will be constant pressure to show upfront benefits for the dollars that are invested in this area. This is where the Enablers step in and provide the Enterprise IT teams with out of the box solutions that can be jump start the whole program.

Infosys HIMI Voice of Customer Analytics Solution

Infosys HIMI Voice of Customer (VoC) Analytics solution is one such Adoption Enabler that has been designed and built to leverage Big Data Technologies and Analytic techniques like Text Mining, NLP and Network analytics to enable Enterprises to improve Customer Experience. The solution provides a holistic approach for processing customer feedback from blogs, customer complaints, discussion forums and call center notes to extract relevant information. Strong text processing capabilities, proprietary content rating algorithms and intuitive tools to analyze unstructured data make this product unique among other similar products in the industry

Solution Overview :

VoC solution provides various listening capabilities with the help connectors and data adapters to popular social media networks such as twitter, Facebook and also to Social media content aggregators. VoC analytical engines leverages Map-Reduce framework for large data analytics. The Solution Overview diagram provides a high-level solution overview of VoC.



Key Features

Strong text processing capability

Infosys HIMI Voice of Customer Analytics solution employs multiple text processing techniques like classification, clustering, key phrase extraction, rule based extraction etc. Techniques for learning from sample data, improving accuracy of classification by incorporating domain specific information are also employed.

Multiple input format support

Input documents in various formats including, but not limited to word, PDF, excel and HTML are supported. Data extraction from online sources is also supported.

Proprietary content rating algorithm

A proprietary content rating algorithm is used to rate the feedback. This algorithm takes into consideration the content, its length, credibility of the websites/ forums and people.

Powerful modeling workbench

Solutions of this nature greatly benefit from domain specific knowledge. The modeling workbench of Infosys HIMI Voice of Customer Analytics solution facilitates easy modeling of domain knowledge. It allows us to capture the domain rules, define and describe the entities and concepts in the domain in a simple manner using an interface.

Out of the box dashboards

Dashboards ensure effective visualization, enable drilling down of data and provide multiple views of data. Graphs, pie charts, dial meter indicators and heat maps are all part of the dashboards.

Business Benefits

Understanding Customer Pain Points: Providing insight into customer sentiments can go a long way into understanding customer concerns. This is instrumental in bringing a customer centric focus to business.

Signify Early Warnings: Customer reviews on product launch and later identifies customer likes and dislikes which influence the product features and play a great role in the success of a product.

Brand / Vendor Comparison: Customer review analysis can be very effective for making a comparison across products and brands. This can give much insight into to competitor businesses and strategies.

Improve Operational Efficiency: It is possible to detect weak links in the services and the operations by correlating insights from customer feedback to transactional, operational and customer demographic data.

Platform – Social Edge for Big data

Social media usage among consumers is growing at a humongous pace resulting in huge amount of data created every minute. The growth in usage of Smart phones', location based apps and more "Internet of Things", the data generated is multiplying at a faster pace. Few statistics on the social data growth:

- More than 250 million tweets are generated in a day and it is increasing at a tremendous speed.
- 30 billion pieces of content shared on Facebook on a monthly basis.
- 40% project growth in global data generated per year.
- Data will grow over 800% in the next 5 years and 80% of these data will be unstructured.

While handling such huge volumes of data poses a significant challenge, at the same time it provides huge opportunities and competitive edge for the enterprises that manages it. The need for a robust platform arises due to the below challenges faced by the enterprises in handling Big Data.

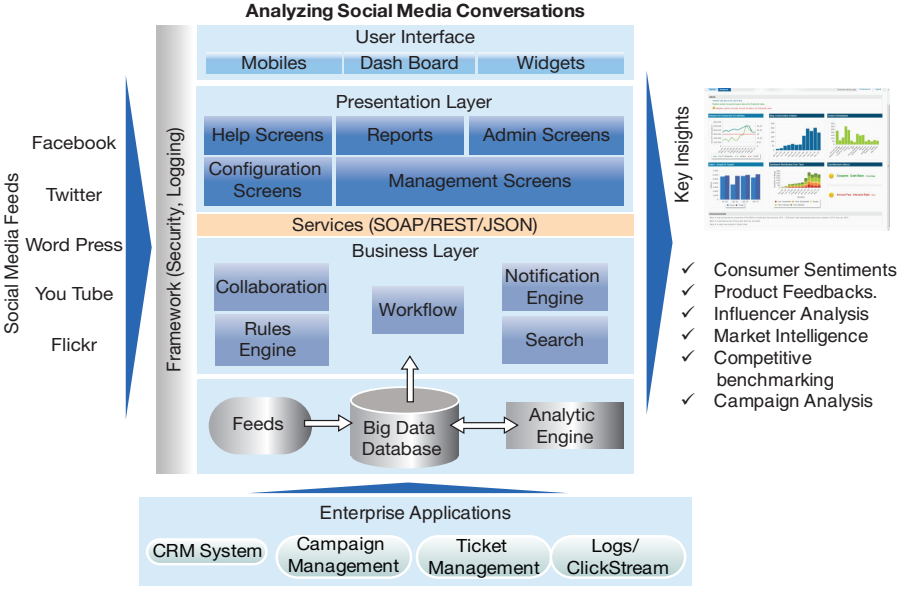
- Setting up a high performance scalable architecture which involves huge investment in terms of hardware and software to set up the infrastructure for processing Big data.
- Aggregating and storing the poly-structured (often unstructured) data from different sources at near real time.
- Getting in not so hygienic external data into the organizations premises, exposing security risks
- Analyzing and making sense of this massive data on an ongoing basis to derive actionable insights.

Infosys SocialEdge – Big Data Analytics Platform

SocialEdge is a comprehensive social media platform which empowers businesses to harness the power of social media to understand customer sentiments, improve branding, deepen engagement and foster innovation. The Big Data Analytics capability of the platform helps in combining social media data streams with enterprise applications in a powerful way to derive meaningful insights on the social conversations. It can also provide analytics on the employee sentiments based on the foot print generated by employees on various corporate systems.

The SocialEdge platform covers the below critical aspects of Big Data processing:

- Data collection and Aggregation: The platform supports data collection and aggregation over diversified data sources across the internet.
  - Over 250 million blogs, 5 million Forums, message boards and usenet groups
  - Micro blogs (Twitter) and Social networking sites (Facebook, Linkedin)
  - Consumer review and shopping sites (Amazon, Bestbuy) , News and mainstream media sites
  - Video and photo sharing sites (Youtube, Flickr) and wikis.



Infosys SocialEdge - BigData Analytics Platform

- Data Processing and Analytics Engine: The platform is driven by a powerful predictive and text analytics engine that can process huge volume of data at a higher speed and identifies sentiment, buzzword, predictive, correlation and influencer data.

- Data Visualization and Reporting: Reporting capability with rich UI interface and ability to create customizable widgets, dashboards and drill down capabilities. It also provides built in engagement capabilities where response for social conversation can be managed and tracked.

Key Differentiators of the platform

- *Advanced Search:* Fast-loading search engine with full Boolean capability and search filters.
- *Dive into the details*: Drill down into data points on a chart to understand what drives changes in volume or sentiment at the post level.
- *Customizable Dashboard*: Enables monitoring of critical reports as widgets on an ongoing basis.
- *Advanced Insights*: Create meaningful brand level insights by combining social and traditional KPIs, consumer sentiments, Influencer Analysis and Competitive Benchmarking.
- *End to End Services*: Platform delivered in enterprise SAAS model with single point accountability, application ownership, robust infrastructure, BPO and consulting services.

Having a scalable Big Data processing platform to analyze social data will help enterprises in,

- Responding faster to a social outburst before negative sentiments go viral in the social world.
- Taking key decisions on products and services based on consumer feedbacks.
- Identifying and engaging key influencers who are impacting the increase or decrease of sales of a particular product.

Client Success Story using SocialEdge – Big Data Analytics

SocialEdge – BigData Analytics platform used by a US high tech major for analyzing social media data.

Business Drivers:

- Understand customer sentiments in social media conversations and improve customer satisfaction, deepen brand loyalty and advocacy during and after the product launch.

Solution Benefits:

- Derive actionable insights for marketing and business managers from a large volume of unstructured social data.
- Ability to monitor issues of customers and pre-empt potential threats by responding proactively.
- Monitor customer sentiment and perception by keeping real time tab on the pulse of customer.
- Anticipate and gauge the product Adoption lifecycle by interpreting the trends in social media.

References

1. Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, May 2011
2. Big Data' Is Only the Beginning of Extreme Information Management by Gartner (Mark A. Beyer, Anne Lapkin, Nicholas Gall, Donald Feinberg, Valentin T. Sribar)
3. Privacy-Preserving Data Mining: Models and Algorithms. Edited by Charu C. Agarwal and Phillip S .U, Kluwer Academic Publishers, 2007.
4. Privacy Preserving Data Mining, Rakesh Agarwal and Ramakrishan Srikanth, IBM Almaden Research Center, 2000.
5. SQL Server 2008 R2 Glossary StreamInsight http://msdn.microsoft.com/en-us/library/ee378962.aspx
6. Demand Response Measurement & Verification - http://www.smartgrid.gov/sites/default/files/pdfs/demand_response.pdf
7. T. Hey, S. Tansley, and K. Tolle, Eds., The Fourth Paradigm: Data-Intensive Scientific Discovery.,2010.
8. A. Wagner, S. Speiser, and A. Harth, "Semantic web technologies for a smart energy grid: Requirements and challenges," in ICWS, 2010. http://iswc2010.semanticweb.org/pdf/506.pdf
9. Hadoop - The definitive guide by Tom White
10. Massive Data Analytics and the Cloud A Revolution in Intelligence Analysis by Booz Allen Hamilton (Michael Farber, Mike Cameron, Christopher Ellis, Josh Sullivan, Ph.D.)

Contributing Authors

Making it Real – Industry Use Cases

Arun Thomas George, *PE;* **Bhushan Dattatraya Masne,** *Infosys Labs;* **Brahma Acharya,** *Cloud;* **Girish Viswanathan,** *RCL;* **Jagdish Bhandarkar,** *E&R;* **Jayanti Vemulapati,** *PE;* **Joseph Alex,** *PE;* **Prem Kumar Karunakaran,** *PE;* **Sajith Abdul Salim,** *PE;* **Sangeetha S,** *E&R;* **Sourav Mazumder,** *Cloud;* **Subramanian Radhakrishnan,** *RCL;* **Sumit Sahota,** *Infosys Labs;* **Venkataramani M,** *Cloud;* **Vinay Prasad,** *FSI;* **Yogesh Bhatt,** *MFG;* **Yuvarani Meiyappan,** *E&R*

Making it Real – Key Challenges

Arun Viswanathan, *Infosys Labs;* **Neminath Hubballi,** *Infosys Labs;* **Rajeshwari Ganesan,** *Infosys Labs;* **Shyam Kumar Doddavula**, *Infosys Labs;* **Soumen Chatterjee,** *Cloud;* **Sudhanshu Hate,** *Infosys Labs*

Making it Real – Infosys Adoption Enablers

Kiran N.G, *E&R;* **Pradeep Kumar M,** *BIZP;* **Rajeev Nayar,** *Cloud;* **Shanmugam Periasamy,** *BIZP;* **Venugopal Subbarao,** *Infosys Labs*

Reviewers

Lakshmanan G, *E&R;* **Siva Vaidyanatha,** *RCL;* **Subrahmanya S.V**, *E&R*

Designers

Chandrashekhar Hegde, *CDG;* **Srinivasan Gopalakrishnan,** *CDG*

Sponsors

Satyendra Kumar, *Senior VP, Head – Quality, Tools and Software Reuse*

Srikantan Moorthy, *Senior VP, Head – Education and Research*

Subrahmanyam Goparaju, *Senior VP, Head – Infosys Labs*

Acknowledgement

Big Data Spectrum themed on 'Making it Real', gives the insights into applying Big Data into real world and discussing the overall way in which any enterprise can embrace Big Data and its related technologies.

This project would not have been possible without the immense support of Kris Gopalakrishnan, Co-Chairman, Infosys Ltd. We are highly grateful to Shibulal S.D, CEO, Infosys Ltd for his encouragement.

We also would like to thank Srikantan Moorthy, Senior VP and Head - Education & Research, Subrahmanyam Goparaju, Senior VP and Head – Infosys Labs, Satyendra Kumar, Senior VP and Head - Quality for their constant support throughout this project.

We would like to thank Dr. Phil Shelley, CTO, Sears Holdings and Doug Cutting, Co-founder, Apache Hadoop for providing their insights in the Q&A section.

We would like to thank Ms. Melissa Hick, Bhava Communications, Rob Lancaster, Cloudera and A. Sriram, Cloud Practice, Infosys Ltd, for their help on Q&A section with Doug Cutting. We are grateful to Mayank Ranjan, RCL, Infosys Ltd, for his help on Q&A section with Phil Shelly.

Rajasimha S, CDG, has contributed to this project. His timely and valuable support in experience design has helped us tremendously. Our sincere thanks to Sarma KVRS, E&R for his help and support on Intellectual Property related details.

Thanks are due to Sanjita Bohidar, Amit Shukla, Satish Kumar Kancherla, Sujith Penta who helped in proofreading.

Infosys® | Building Tomorrow's Enterprise

For more information, contact askus@infosys.com

About Infosys

Many of the world's most successful organizations rely on Infosys to deliver measurable business value. Infosys provides business consulting, technology, engineering and outsourcing services to help clients in over 30 countries build tomorrow's enterprise.

For more information about Infosys (NASDAQ:INFY), visit www.infosys.com.