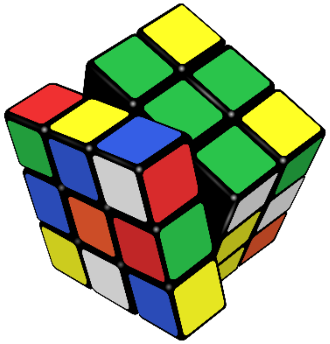


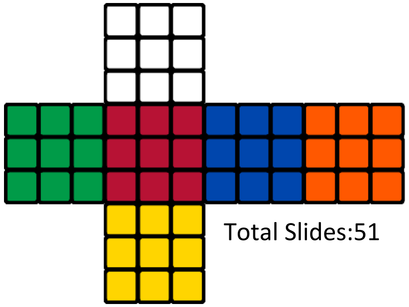
Big Data Engineering For Analytics

Introduction to Big Data Engineering

Dr Venkat Ramanathan

(rvenkat@nus.edu.sg)

NUS-ISS



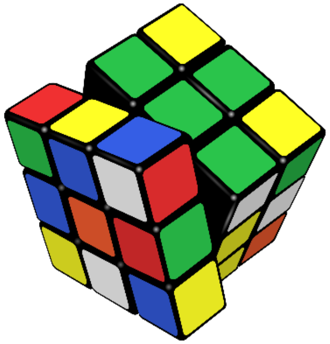
Total Slides:51

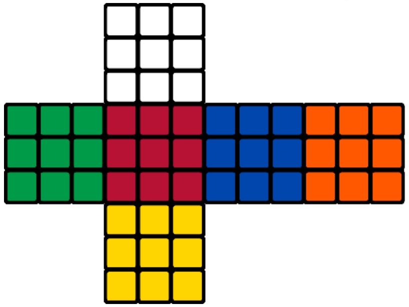
Learning Objectives

- Analyze, classify and characterize the **“Big Data”**
- Analyze critically the **design challenges** in building a distributed computing platform.
- Appraise the needful ***components*** by analyzing key *characteristic* requirements as demanded by the distributed system being built.
- Evaluate the suitability of **complementary ways** in designing distributed computing architecture.
- Analyze, interpret and use appropriate additional models for aspects such as communication interaction, failure and security.

Agenda

- Introduction to Big Data
- Design Challenges in Building Big Data Engineered Solutions.
- Components of Big Data Engineered Solutions.
- Use Cases for Data Science
- Summary





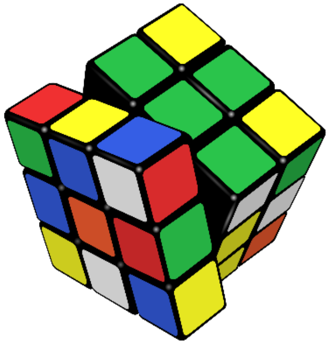
Introduction to Big Data

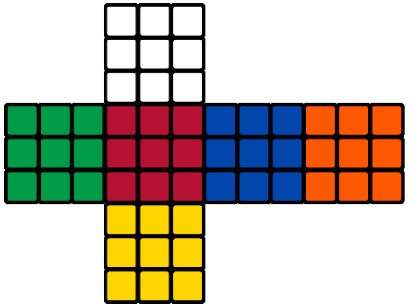
There sat that beautiful big machine whose sole job was to copy things and do addition. Why not make the computer do it? That's why I sat down and wrote the first compiler. It was very stupid. What I did was watch myself put together a program and make the computer do what I did.

Admiral Grace Hopper

The Course… BEAD

BIG DATA ENGINEERING FOR ANALYTICS



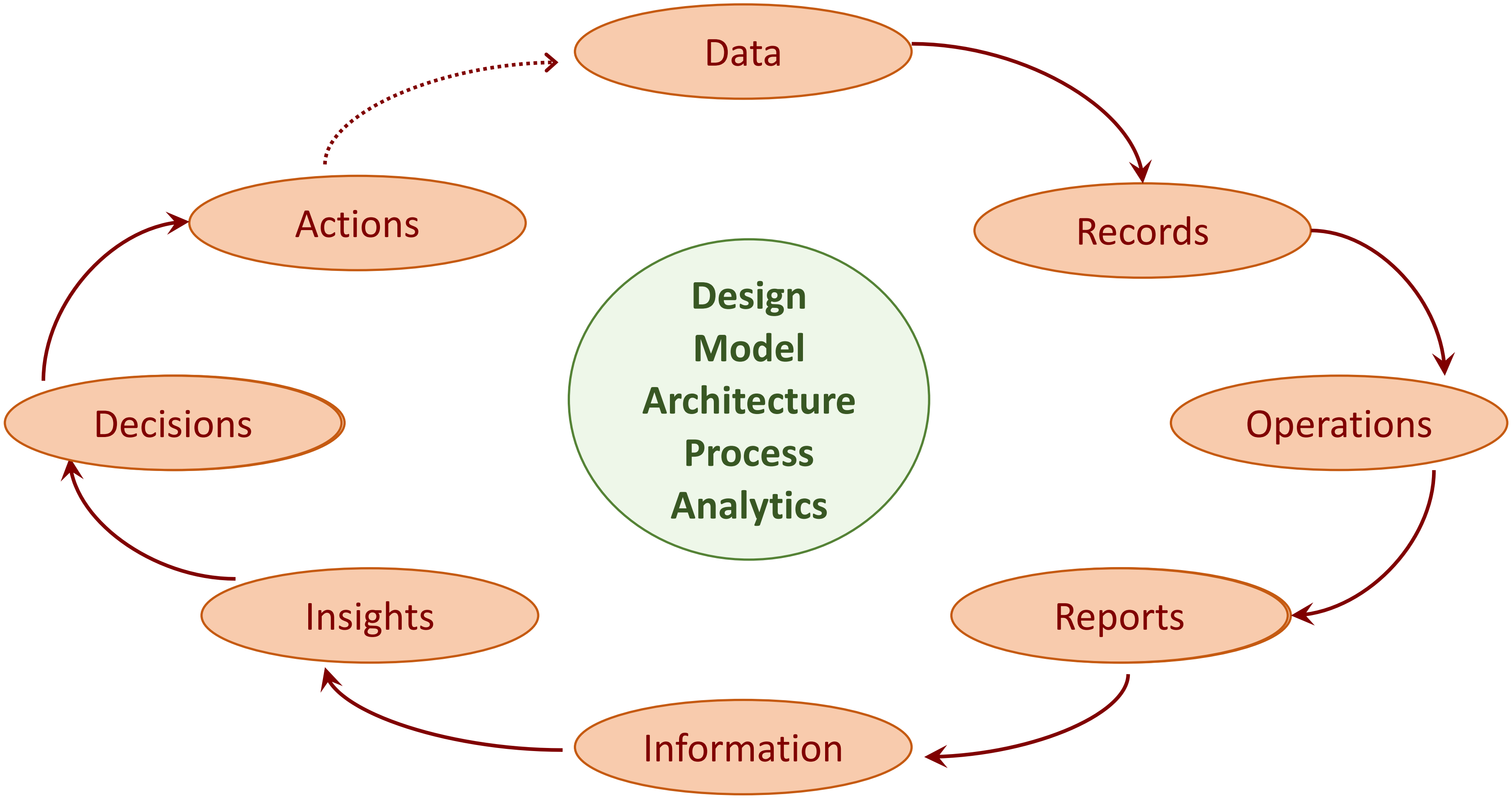


DATA:

What is data?

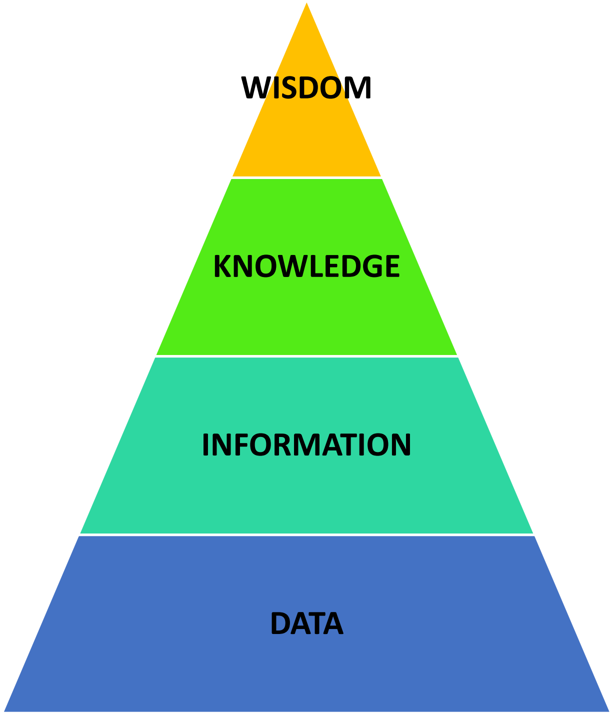
Why do we need data?

Need for sourcing & storing data



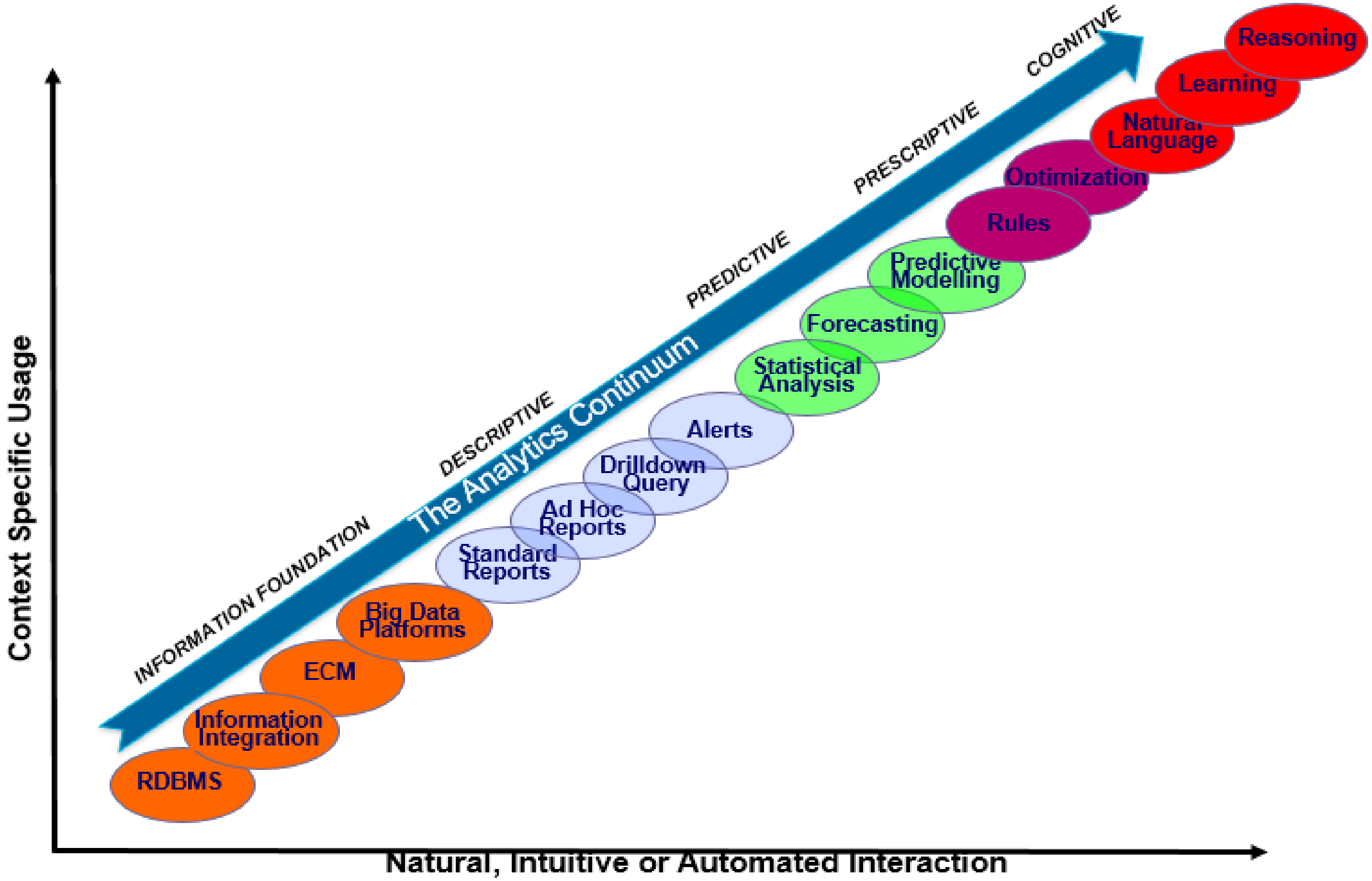
Analytics

- What is Information (step towards Analytics)?
- Derivative from DIKW Model

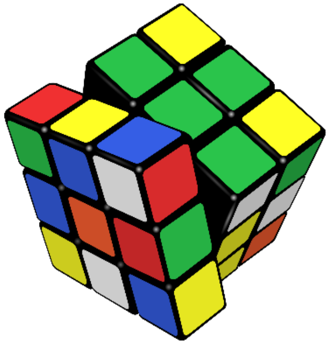


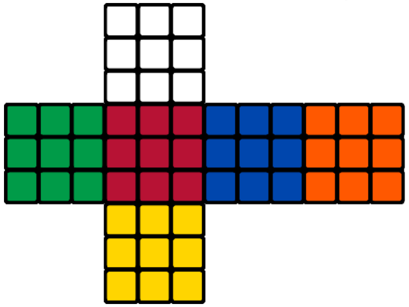
Applied	•No Rain, cloudy so outdoor activities ok •Moderate haze, so restrict to healthy participants and no physically strenuous games.
Context	• Annual Day event (out door and indoor activity)
Meaning	• Singapore Island, West, Time 13:01:05PM, PSI 210 PM2.5 μg/m^3 , Temperature 32^{0}C, Humidity 62%, Cloudy
Raw	• SG-W:13:01:05:210:32:62:C

The Analytics Continuum



Source: https://www.ibm.com/developerworks/community/blogs/jfp/entry/Cognitive_Computing_vs_Analytics?lang=en

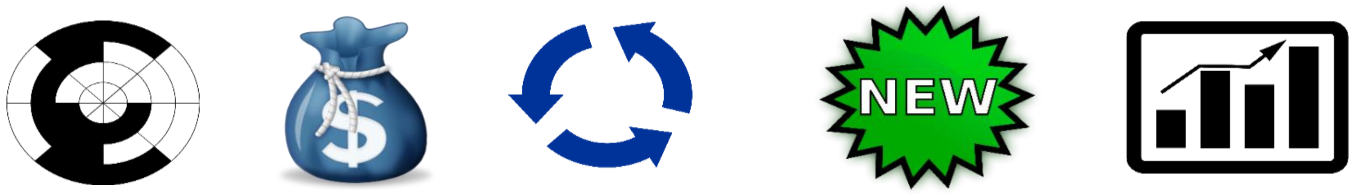




Business Analytics

Business Drivers for Analytics

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)



OPTIMIZE MONITIZE TRANSFORM NEW BUSINESS GROWTH

Business Questions

- Bank (eg: a large bank such as DBS)
  - Loans
    - Who should I campaign to offer loan?
    - Will this customer default his loan?
    - What kind of financial activities does the customer involve in?
  - Credit Card
    - Risk management – eg: Is the current transaction a fraudulent usage?
    - Usage patterns for campaigns, eg: What products should we sell to this client?
  - To answer the above first we determine what or wherefrom the data is sourced?
    - Existing records with banks (eg: Transaction records in savings, credit & other ac)
    - Point of sales systems where cards are used.
    - Other external systems like credit rating organisations, Govt. organisations etc.
    - Profiling of customer based on wealth distribution & distribution.

Business Questions

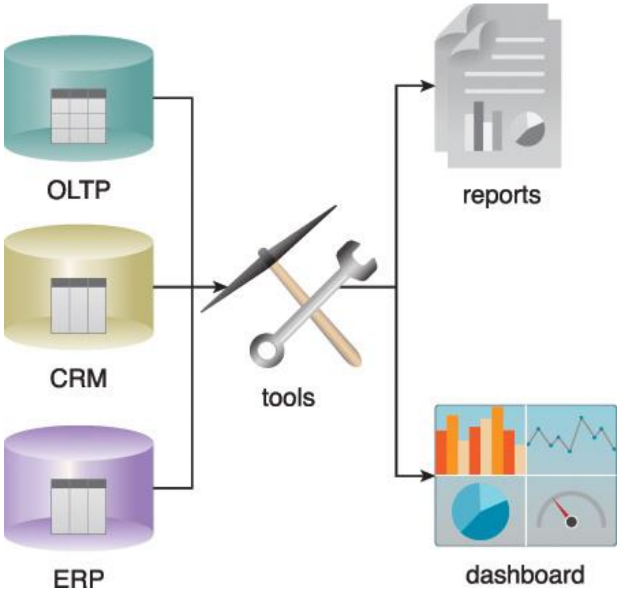
- How can we answer those questions?
  - Need Data
  - Need to analyse data
  - Need to infer data
  - Need to predict outcomes
- Using data for decisions always existed; *what is new?*
  - Look at the cases described
    - Large amounts of Data
    - Data is from heterogamous sources and heterogeneous types
    - Data is growing rapidly
    - etc…

1. Descriptive Analytics

Descriptive analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.

Examples:

- What was the sales volume over the past 12 months?
- What is the number of support calls received as categorized by severity and geographic location?
- What is the monthly commission earned by each sales agent?



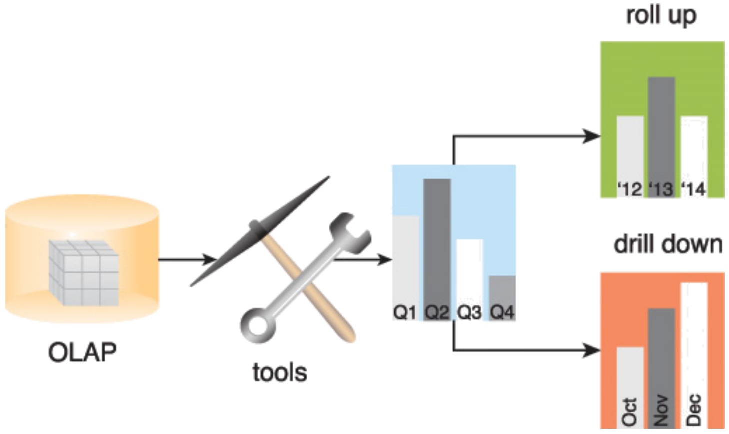
2. Diagnostic Analytics

Diagnostic analytics aims to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.

- The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

Examples:

- Why were Q2 sales less than Q1 sales?
- Why have there been more support calls originating from the Eastern region than from the Western region?
- Why was there an increase in patient re-admission rates over the past three months?



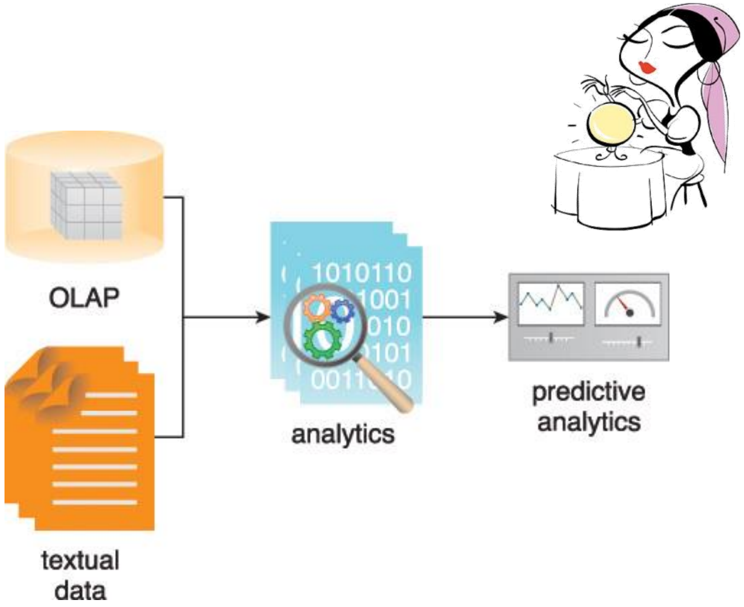
3. Predictive Analytics

Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. With predictive analytics, information is enhanced with meaning to generate knowledge that conveys how that information is related.

- Models use strength and magnitude of the associations (of past events).
- Models depends on the conditions under which the past events occurred which should be changed if the conditions change.

Examples:

- What are the chances that a customer will default on a loan if they have missed a monthly payment?
- What will be the patient survival rate if Drug B is administered instead of Drug A?
- If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?



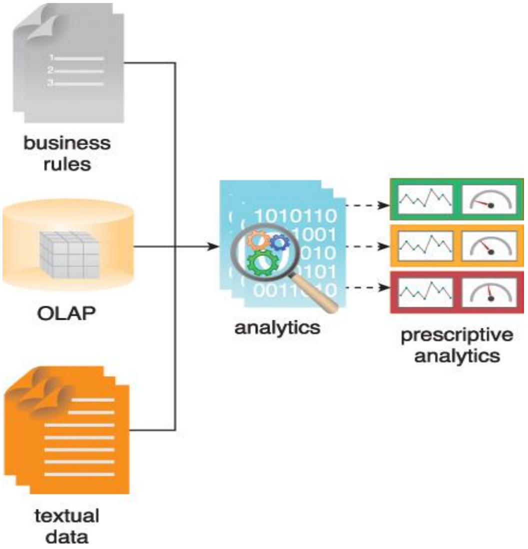
4. Prescriptive Analytics

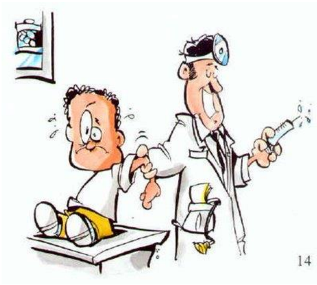
Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why.

- Provide results that can be reasoned about because they embed elements of situational understanding that can be used to gain an advantage or mitigate a risk.
- Provides more value than any other type of analytics and correspondingly require the most advanced skillset, specialized software and tools.
- Computes various outcomes and suggests the best course of action for each each.

Examples:

- Among three drugs, which one provides the best results?
- When is the best time to trade a particular stock

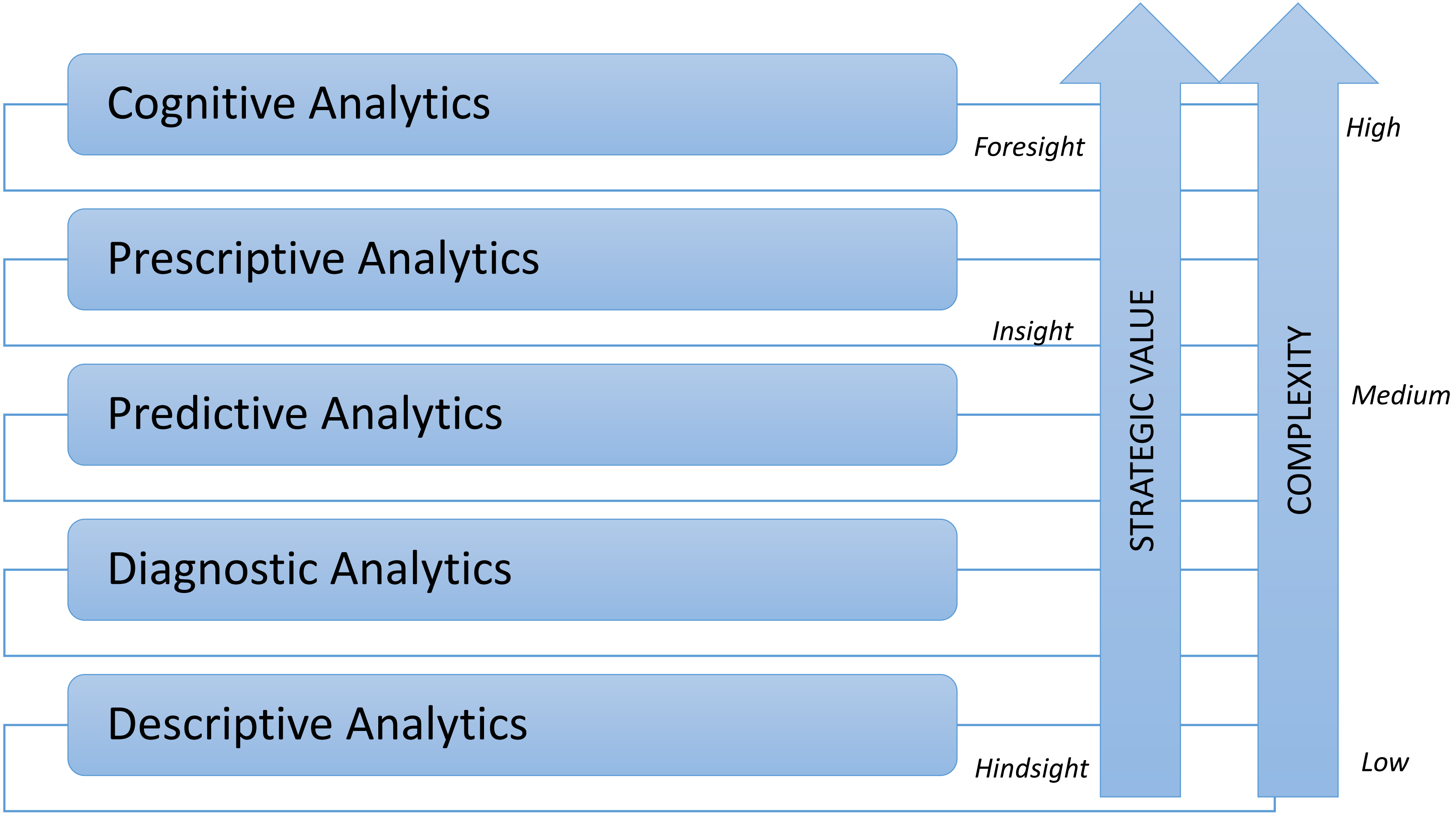




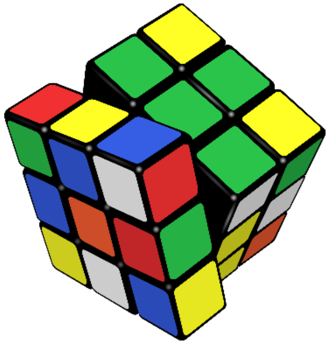
5. Cognitive Analytics

- Cognitive computing combines artificial intelligence and machine-learning algorithms, in an approach which attempts to reproduce the behavior of the human brain.
- Cognitive analytics applies intelligent technologies to bring all of these data sources within reach of analytics processes for decision making and business intelligence.
  - ➢ Cognitive analytics is a data forward approach that starts and ends with what's contained in information. This unique way of approaching the entirety of information (all types and at any scale) reveals connections, patterns and collocations that enable unprecedented, even unexpected insight.
  - ➢ A cognitive system can provide real-time answers to questions posed in natural language by searching through massive amounts of information that have been entered into its knowledge base, making sense of context, and computing the most likely answer. As developers and users “train” the system, answers become more reliable and increasingly precise over time.

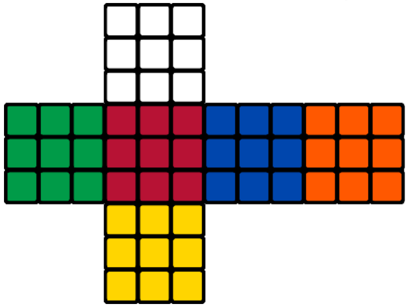
Categories of Analytics



Inspiration: Adapted & Enhanced from Big Data Fundamentals: Concepts, Drivers & Techniques by Thomas Erl et al.



Big Data Engineering For Analytics



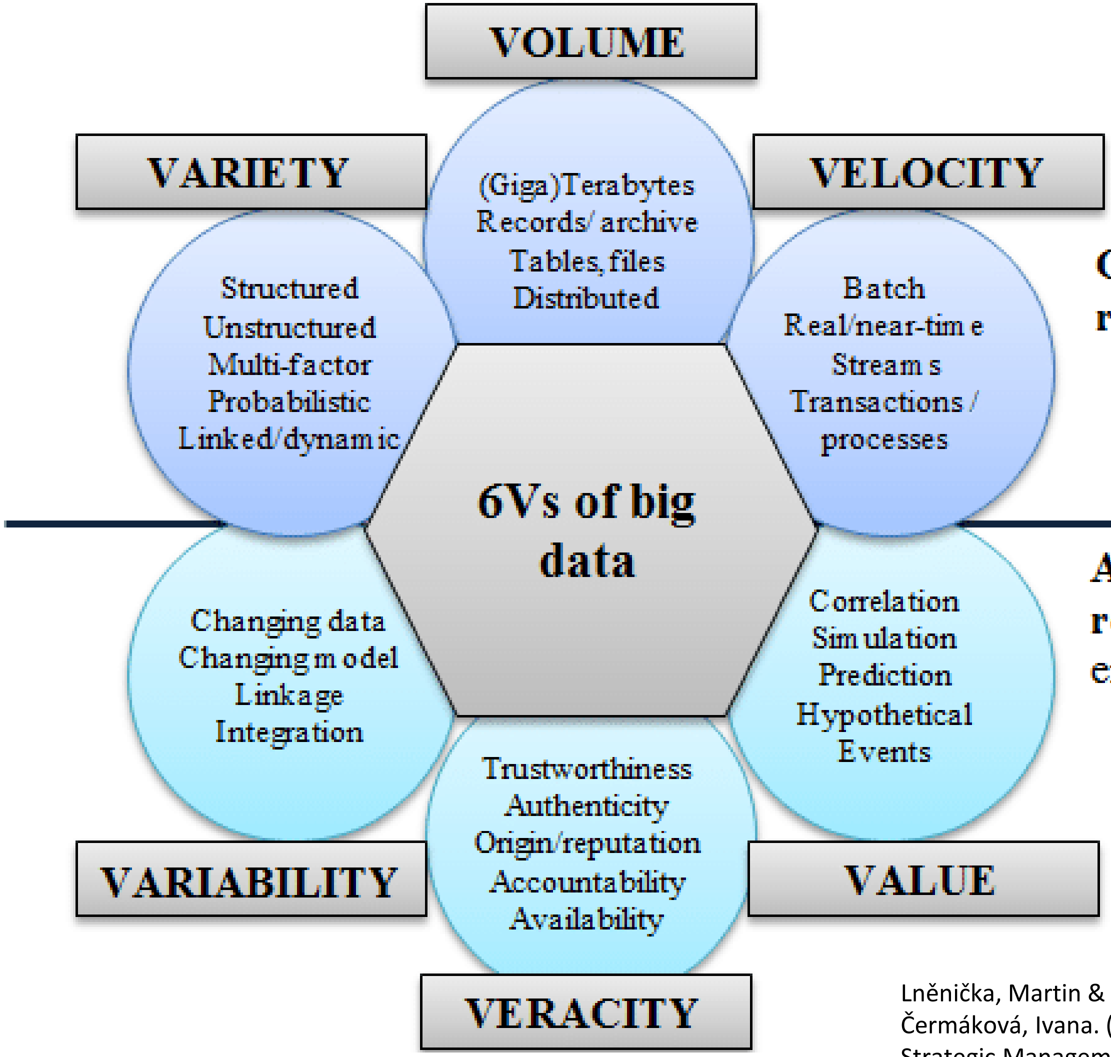
Big Data

What is Big Data

- Another Buzz Word?
- Precise Decision Enabler?
- New Technology?

Modern Day Data Characteristics – Big Data

- ***Big Data*** are large corpus of datasets with characteristics such as volume, variety, velocity, veracity, variability and value. It requires a scalable architecture for efficient storage, manipulation, and analysis.
  - **Volume** refers to the vast amount of data generated every second.
  - **Variety** refers to the different types of data we can now use.
  - **Velocity** refers to the speed at which new data is generated and the speed at which data moves around.
  - **Veracity** refers to the uncertainty or trustworthiness of the data.
  - **Variability** refers to the change in data meaning/models over time.
  - **Value** refers to our ability turn our data into value.
- ***Big Data Engineering*** deals with advanced techniques that harness independent resources for building scalable data systems when the above mentioned characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.



6Vs of big data

VOLUME
(Giga)Terabytes
Records/ archive
Tables, files
Distributed

VARIETY
Structured
Unstructured
Multi-factor
Probabilistic
Linked/dynamic

VELOCITY
Batch
Real/near-time
Streams
Transactions /
processes

VARIABILITY
Changing data
Changing model
Linkage
Integration

VERACITY
Trustworthiness
Authenticity
Origin/reputation
Accountability
Availability

VALUE
Correlation
Simulation
Prediction
Hypothetical
Events

Generic big data requirements:

- variety,
- volume,
- velocity.

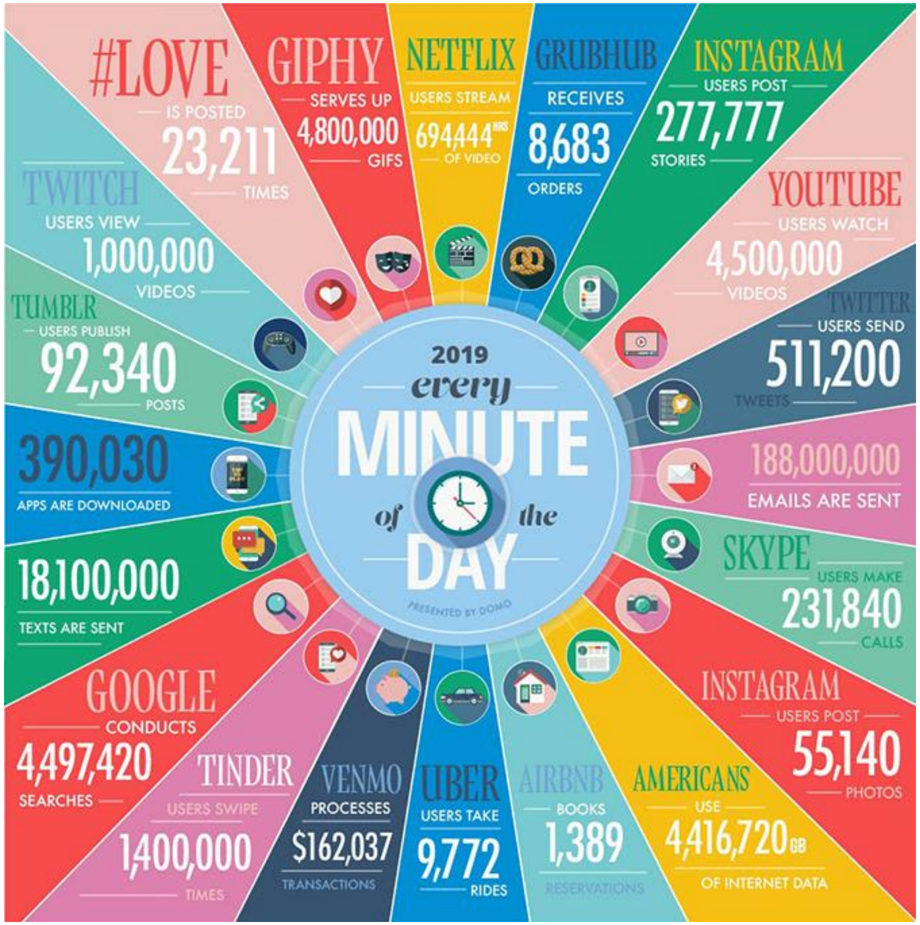
Acquired big data requirements (after entering system):

- variability,
- veracity,
- value.

Lněnička, Martin & Máchová, Renáta & Komárková, Jitka & Čermáková, Ivana. (2017). Components of Big Data Analytics for Strategic Management of Enterprise Architecture. .

Let's contextualise the V's with a case scenario – eg: Google Maps!

Data Never Sleeps….. !!!!



Data is constantly pouring out of our smartphones, smart watches, smart TVs, and countless other devices that are all connected—and it continues to proliferate at an astounding rate. But just how much data is generated every minute in 2019?

Reference: https://www.domo.com/blog/data-never-sleeps-6/

Data can take different forms.

- ***Structured data*** refers to data that has a defined length and format
  - ➢ Computer- or machine-generated: Machine-generated data generally refers to data that is created by a machine without human intervention.
  - ➢ Human-generated: This is data that humans, in interaction with computers, supply.
- ***Unstructured data*** is data that does not follow a specified format
- ***Semi-structured data*** refers to data that falls between structured and unstructured data.
  - ➢ Semi-structured data does not necessarily conform to a fixed schema (that is, structure) but can be interpreted (mainly by self-describing data like having simple label/value pairs).

More Details in the subsequent Data Design Session

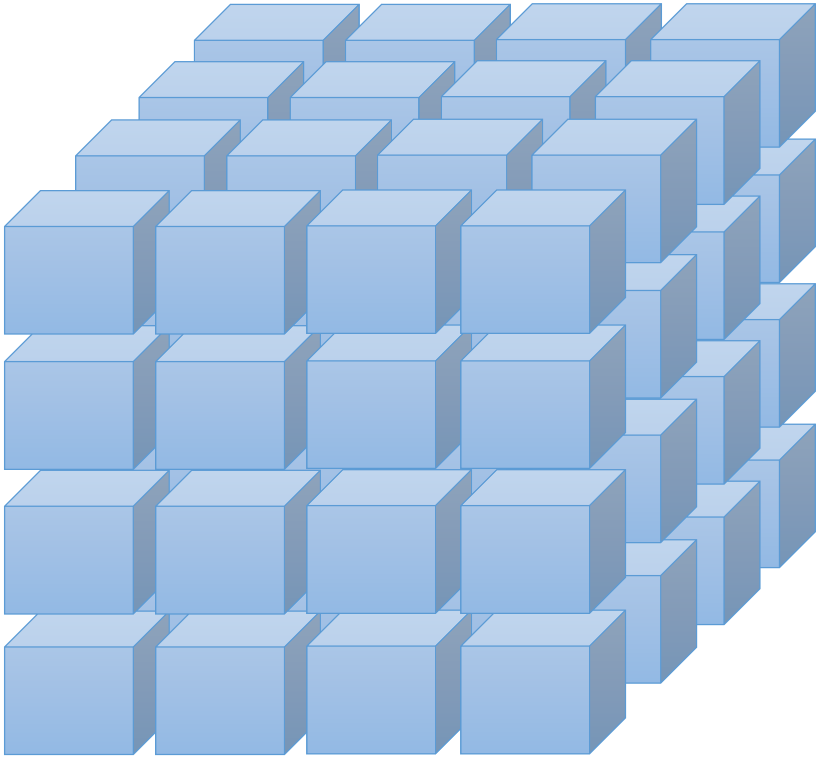
Data is either moving or at rest . . .

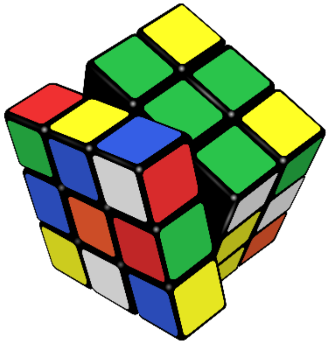
- There are three basic states of data: data at rest, data in motion, and data in use.
  - **Data at rest**
    - Data at rest is a term that refers to data stored on a device or backup medium in any form.
    - It can be data stored on hard drives, backup tapes, in offsite cloud backup, or even on mobile devices.
    - Data at rest is a term that refers to data stored on a device or backup medium in any form. It can be data stored on hard drives, backup tapes, in offsite cloud backup, or even on mobile devices.
  - **Data in motion**
    - Data in motion is data that is currently traveling across a network or sitting in a computer's RAM ready to be read, updated, or processed.
    - This data in motion (usually encrypted) includes data moving across a cables and wireless transmission. It can be emails or files transferred over FTP or SSH.
  - **Data in use**
    - Data in use is data that is not just being stored passively on a hard drive or external storage media. This is data that is being processed by one or more applications.
    - This is data currently in the process of being generated, updated, appended, or erased.

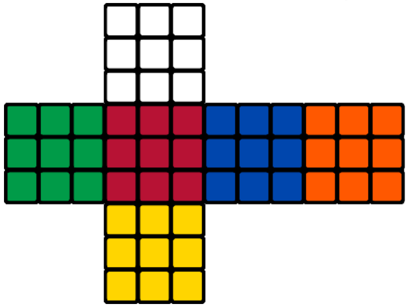
Reference: http://aspg.com/three-states-digital-data/
Three states of Data https://en.wikipedia.org/wiki/Digital_data

Big Data Defined

Big Data defines a situation in which data sets have grown to such enormous sizes and consists of a variety of structures and are obtained from, stored at or available in various locations that conventional information technologies can no longer effectively handle either the size of the data set or the scale and growth of the data set.







Meeting the Technology Challenge posed by Big Data

Big Data Processing – Options and Approaches

- Can existing Platform, Architecture, Tools and Technologies handle this?
  - Option 1: Scale Up
  - Option 2: Scale Out

Big Data Processing – Options and Approaches

- Merits of Scale up?
  - ➢ Well we live in existing computing models, methods and software

- Demerits
  - ➢ Cost
  - ➢ There is a limit to which we can scale up
  - ➢ Reliability
  - ➢ Performance

We seem to Gravitate to Scale Out option!

But…. What does that mean? A whole new way of Computing!

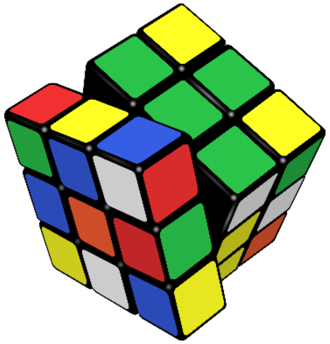
Big Data Processing – Options and Approaches

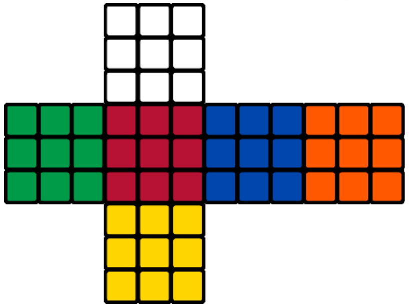
- Scale out essentially means Distributed Systems
- Technological Challenges
  - To process large data sets in multiple systems
    - Well known RDBMS do not meet the need.
  - Nodes may fail – hence we need to have an approach to handle that
  - Number of nodes may not be a constant – we need to cater to that
  - Communication models between nodes need to be established.
    - For example data for same data set or query may be stored across multiple nodes.
    - We need to aggregate it (advantage is we may be able to parallelize)

So what do we need!

A new Distributed Storage and Analysis Technology –

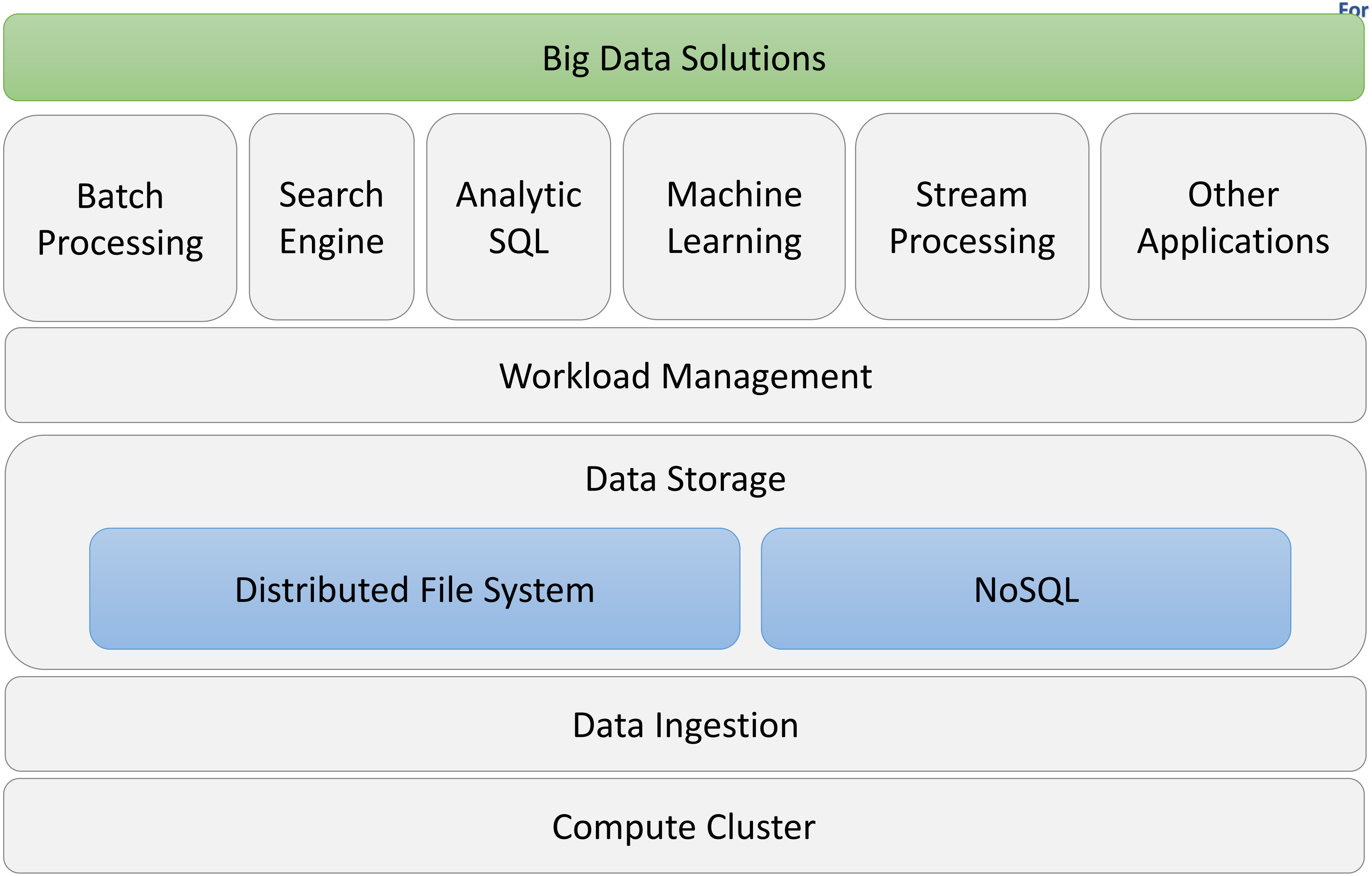
Infrastructure such as Cloud Platform. A Processing Framework like Hadoop!





Implementing the Big Data Solution

Big Data Engineering for Analytics



Big Data Solutions

Batch Processing

Search Engine

Analytic SQL

Machine Learning

Stream Processing

Other Applications

Workload Management

Data Storage

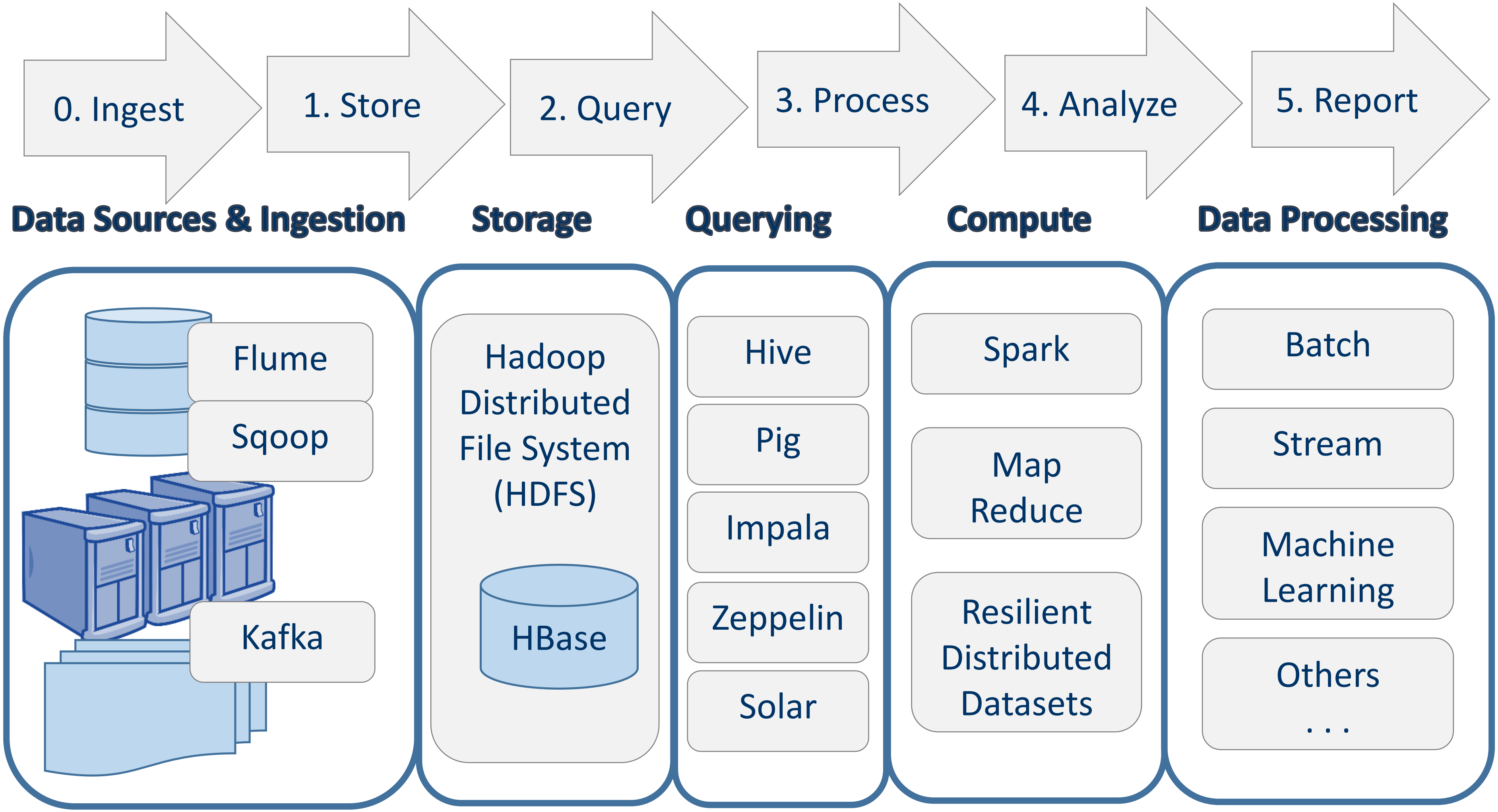
Distributed File System

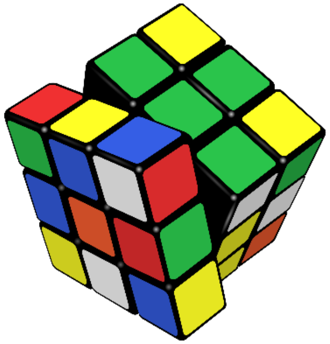
NoSQL

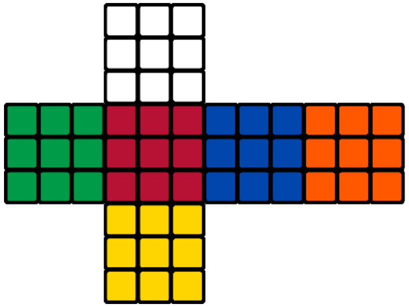
Data Ingestion

Compute Cluster

Processing the Big Data – Technologies







Summary

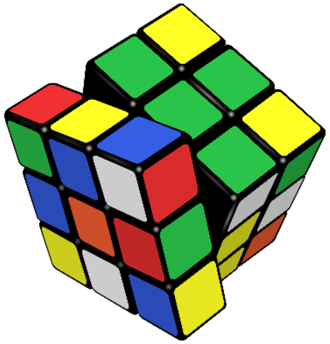
The most dangerous phrase in the language is, ‘We’ve always done it this way.’

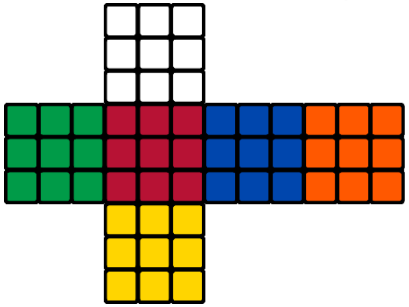
- Admiral Grace Hopper

Key Points

- Autonomous computers that work together to give the appearance of a single coherent system:
  - Resource sharing is the main motivation for constructing distributed systems.
  - Architectural styles reflect way in which both software and hardware are organized to communicate in a distributed system.

- Big Data analysis blends traditional statistical data analysis approaches with computational ones.
  - The overall goal of data analysis is to support better decision-making.
  - Carrying out data analysis helps establish patterns and relationships among the data being analysed.





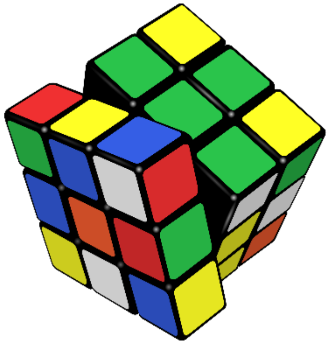
References

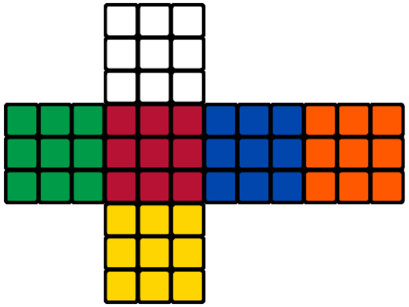
We're flooding people with information. We need to feed it through a processor. A human must turn information into intelligence or knowledge. We've tended to forget that no computer will ever ask a new question.

- Admiral Mdm. Grace Hopper

References

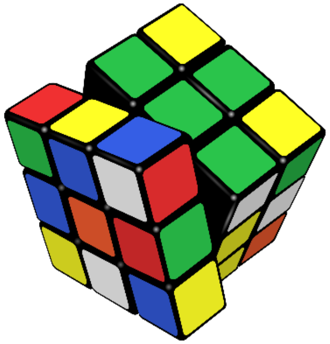
- Coulouris, George F., Jean Dollimore, and Tim Kindberg. *Distributed systems: concepts and design*. pearson education, 2005.
- Tanenbaum, Andrew S. *Distributed operating systems*. Pearson Education India, 1995.
- Verissimo, Paulo, and Luis Rodrigues. *Distributed systems for system architects*. Vol. 1. Springer Science & Business Media, 2012.
- Data Science with Hadoop and Spark: Designing and Building Effective Analytics at Scale, by Casey Stella et al, Published by Addison-Wesley Professional, 2016
- Enterprise Big Data Engineering, Analytics, and Management, by Thomas Roth-Berghofer, Samia Oussena, Martin Atzmueller, Publisher: IGI Global, Release Date: June 2016
- Magazines, newspapers, blogs, Wikipedia, websites and other online resources.



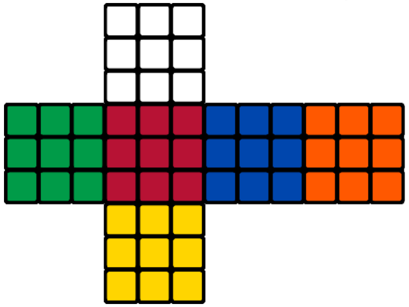


Appendix:

Non-Lecture slides (for reference)



Big Data Engineering For Analytics



Use Cases Big Data Analytics

How many millionaires do you know who have become wealthy by investing in savings accounts? I rest my case.

~Robert G. Allen

Product Recommendation

- Recommender systems have become rather common for online retailers and many other businesses with an online retail presence.

- We are familiar with various flavours of product recommendation techniques used by companies such as Amazon, Netflix, Facebook, LinkedIn, and, more recently, Google/YouTube.

Customer Churn Analysis

- It is well known that keeping an existing customer is often much cheaper than finding a new one.
  - ➢ Whether you are a bank, a retailer, a gaming company, an Internet service provider, a cell phone provider, an airline, or an insurance company, there is a strong desire in almost any business to actively pursue programs for customer retention and prevent customer churn (also known as "attrition").
  - ➢ Churn models differ by industry, due to the different business models and specific customer engagement and lifetime value models.
- Customer churn analysis uses machine learning to predict the likelihood of each customer "leaving."
- Businesses then use this data to drive and guide customer-retention programs (such as discounts or other incentive programs) to encourage these at-risk customers to stay.

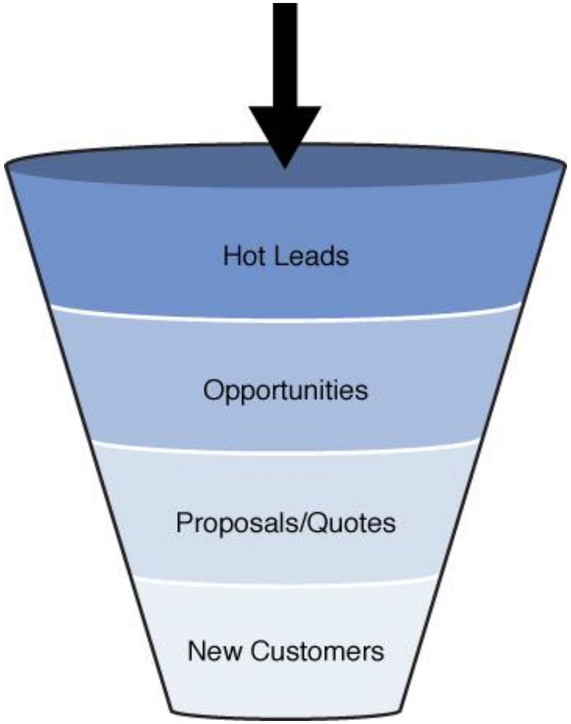
Customer Segmentation

- Customer segmentation is a common technique used to identify segments of customers that behave similarly with regard to their interaction with the business.
  - ➢ A grocery store may be interested in segmenting its customers by the type of food products they purchase. For example, one segment of customers might be “people who favour meat,” while another might be “people who favour gourmet products.”
  - ➢ Similarly, airlines and hotels are interested in segmenting customers into business travellers versus non-business travellers. Airlines are also interested in “domestic passengers” versus “international passengers.”
- An immediate benefit of such segmentation is the ability to increase marketing efficiency. For example, airlines may customize email campaigns based on effective segmentation to achieve much higher response rates.

Sales Leads Prioritization

- Many sales professionals enjoy a pipeline of sales leads that result from good, effective marketing

There are many ways by which a business can prioritize the sales efforts, but one of the most natural parameters is “likelihood to close within N days.”
Applying data science, businesses can model each lead with various features (such as geographic location, customer type, website engagement, previous sales, etc.), and build a predictive model to determine the likelihood of each lead to close within the desired time period.
Based on such models, sales operations improve in efficiency and overall revenues increase.



Sentiment Analysis

- Sentiment analysis is an application of text analytics and natural language processing techniques, with the goal of understanding customer sentiment about a certain topic (e.g., a product or service).
  - ➢ The Web has become an excellent source for assembling consumer opinions. There are now several Web sites containing such opinions, e.g., customer reviews of products, forums, discussion groups, and blogs.
  - ➢ With the increased adoption of crowd-sourced feedback from customers in online forums and the growth of social networks such as Facebook and Twitter, there is a lot of information available about customer sentiment.

Fraud Detection

- Fraud or payment abuse is a serious problem for many businesses as well as government organizations. Every time money changes hands based on some criteria or set of rules, there is potential for fraud and abuse for monetary gain by malicious actors.

- Clearly fraud detection is a critical capability for companies involved in payments such as banks, PayPal, or Square. But fraud detection is also highly effective in improving the bottom line for insurance companies, retailers, and many others.

Predictive Maintenance

- Equipment doesn’t operate forever and will ultimately fail at some point; it always fails sometime in the future.
  - ➢ Unfortunately, such failure may have dire consequences given the binary nature of the failure. There are many examples of this, in various industries. Let’s look at a few:
    - When a component in a cell tower fails, the cell tower may stop operating and many cell phone users in the nearby vicinity may not be able to use their mobile services until the component is fixed and the tower is fully functional again.
    - When an A/C compressor fails in an office building, the employees working there may suffer from poor working conditions for a day or two until a technician is able to fix the problem.
    - If an engine in a helicopter or airplane fails, there could be dire consequences indeed. Fortunately, this is not a common safety issue as testing of these engines is rather thorough before lift-off. Nevertheless, if a failure is found on the ground, the vessel may need lengthy repair, which may result in flight delay or cancellation.
    - If a freezer in a fast-food restaurant fails, the restaurant will need to replace it. However, that may take a few days. What will it do in the meantime with all the frozen food kept in that freezer?

Market Basket Analysis

- A common use case for retailers is known as market basket analysis (also known as affinity analysis or association mining).
  - ➢ In this type of analysis, we are trying to understand the purchasing behavior of the user. More specifically, with market basket analysis, retailers hope to gain insights into which products tend to be purchased together.
  - ➢ Market basket analysis often drives store layout design, where items with strong association are placed strategically close to each other, making it more likely that the customer will purchase the related item.
  - ➢ Retailers can also use the results of market basket analysis for effective marketing campaigns to drive foot traffic into a physical store.

Predictive Medical Diagnosis

- Making decisions about a medical diagnosis is hard, partly because there is often a lot of uncertainty and not enough data.
  - ➢ Furthermore, the implications of an incorrect decision can be dire.
  - ➢ Arming medical professionals with computer-assisted, data-based medical diagnosis tools is a tremendous opportunity to improve healthcare. Let’s explore a few specific applications:
    - ▪ Machine learning algorithms can detect unknown patterns of diagnosis and, if validated clinically, can add to the existing repository of medical knowledge.
    - ▪ Electronic patient records use the ICD10 standard to record existing diagnoses for patients. Sometimes the electronic record is missing a few key diagnoses, and automated diagnosis of disease can be used for identifying these coding gaps.
    - ▪ Various care quality measures such as HEDIS (Healthcare Effectiveness Data and Information Set) can be improved by, for example, screening based on automated diagnosis results.

Insurance Risk Analysis

- Insurance is a risk-based industry. Products such as property, auto, or life insurance are always priced based on risk assessment and the application of the risk pooling principle.
  - ➢ Insurance companies have been using predictive risk modeling for some time, modeling risk based on key indicators such as age, gender, geographic location, and historical data about the consumer.
    - ▪ For example, it is well known that younger drivers tend to be more accident-prone than experienced drivers. Thus, auto insurance companies will typically charge a higher premium for drivers under the age of 25.
  - ➢ Since accurate risk analysis is so critical to the profitability of an insurance company, every attempt is made to improve this and gain a competitive edge.
    - ▪ For example, auto insurance companies are looking at sensor data coming from automobiles (GPS data, etc.) as a new source of data to be used to improve accuracy of risk prediction. By tracking driving behaviour, the insurance company can more accurately assess the risk of accident for that driver.

Predicting Oil and Gas Well Production Levels

- The basic asset of any oil and gas company is the well, from whence oil and natural gas are produced.
  - ➢ Oil and gas companies such as Schlumberger, Haliburton, Noble Energy, and Chesapeake therefore invest heavily in research and development to maximize oil production levels, resulting in direct impact on the top line of the business.
  - ➢ There are many variables that may impact the production levels of a given well.
  - ➢ With sensor data, geophysics data about the well, and other data sources, models can be constructed to predict well production.
  - ➢ With this predictive model, the oil and gas company can understand what impacts production levels and address issues negatively impacting this level ahead of time, resulting in streamlined production and increased revenues.