KTU NOTES

The learning companion.

KTU STUDY MATERIALS | SYLLABUS | LIVE NOTIFICATIONS | SOLVED QUESTION PAPERS

Website: www.ktunotes.in

CST322 – Data Analytics

Module – IV

Compiled by,
Mr. Shyam Krishna K
Asst. Professor, Dept. of CSE
Sahrdaya CET

Syllabus

Module - 4 (Big Data Analytics)

Big Data Overview – State of the practice in analytics, Example Applications - Credit Risk Modeling, Business Process Analytics.

Big Data Analytics using Map Reduce and Apache Hadoop, Developing and Executing a Hadoop Map Reduce Program.

Big Data Overview

Big Data Overview

- Data is created constantly, and at an ever-increasing rate
- Devices and sensors automatically generate diagnostic information that needs to be stored and processed in real time.
- Merely keeping up with this huge influx of data is difficult, but substantially more challenging is analyzing vast amounts of it, especially when it does not conform to traditional notions of data structure, to identify meaningful patterns and extract useful information.

Big Data Overview

- Several industries have led the way in developing their ability to gather and exploit data:
  - Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy
  - Mobile phone companies analyze subscribers' calling patterns
  - For companies such as LinkedIn and Facebook, data itself is their primary product.

Big Data Overview

- Three attributes stand out as defining Big Data characteristics:
  - Huge **volume** of data
  - Complexity of data types and structures - **Variety**
  - Speed of new data creation and growth – **Velocity**

- Due to its size or structure, Big Data cannot be efficiently analyzed using only traditional databases or methods.

- Big Data problems require new tools and technologies to store, manage, and realize the business benefit.

Definition

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value

implies that organizations will need new data architectures and analytic sandboxes, new tools, new analytical methods, and an integration of multiple skills into the new role of the data scientist,

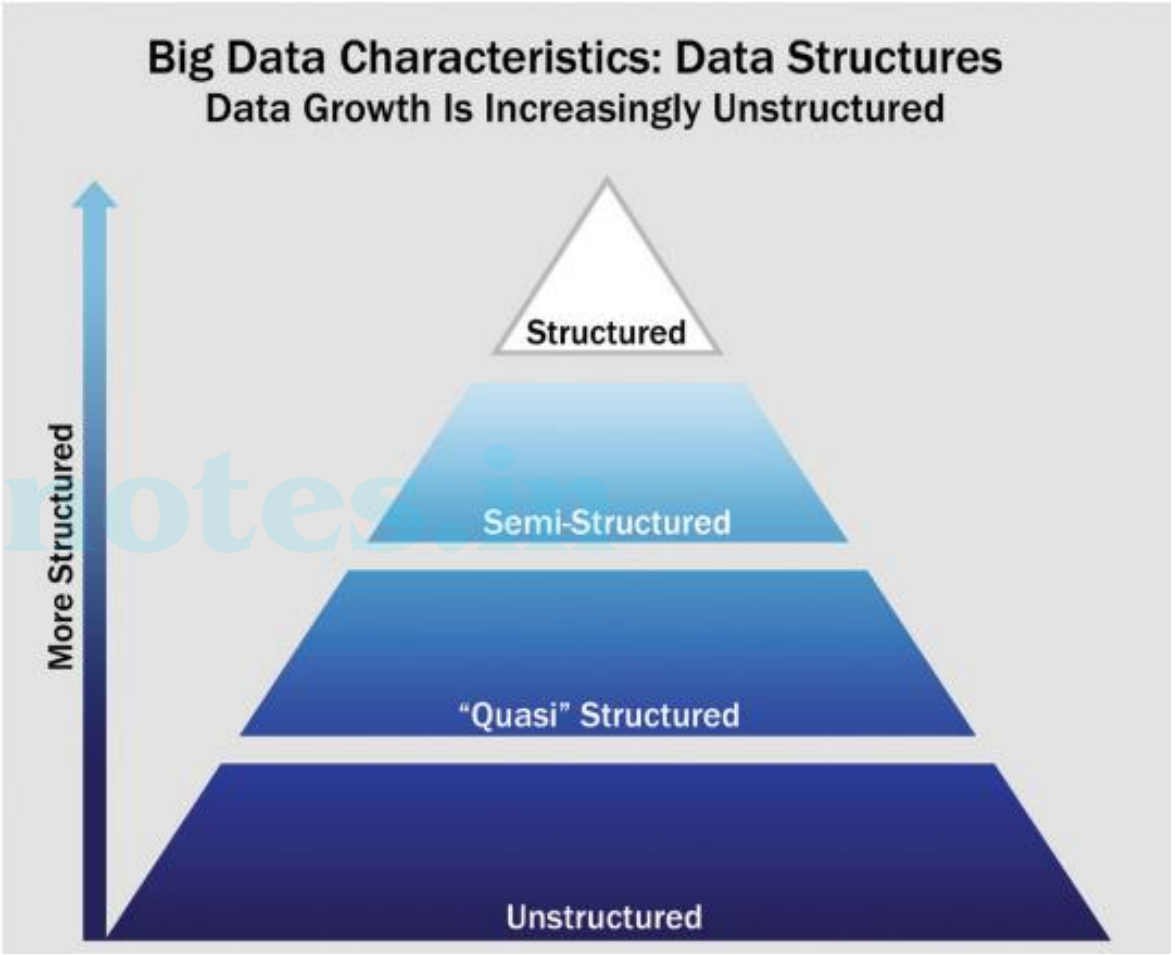
What's Driving Data Deluge?

Data Structures

- Big data can come in multiple forms, including structured and non-structured data such as financial data, text files, multimedia files, and genetic mappings.
- Contrary to much of the traditional data analysis performed by organizations, most of the Big Data is unstructured or semi-structured in nature, which requires different techniques and tools to process and analyze.
- Distributed computing environments and massively parallel processing (MPP) architectures that enable parallelized data ingest and analysis are the preferred approach to process such complex data
- About 80-90% of future data growth is coming from non-structured data types

Data Structures

Although analyzing structured data tends to be the most familiar technique, a different technique is required to meet the challenges to analyze semi-structured data (shown as XML), quasi-structured (shown as a clickstream), and unstructured data.



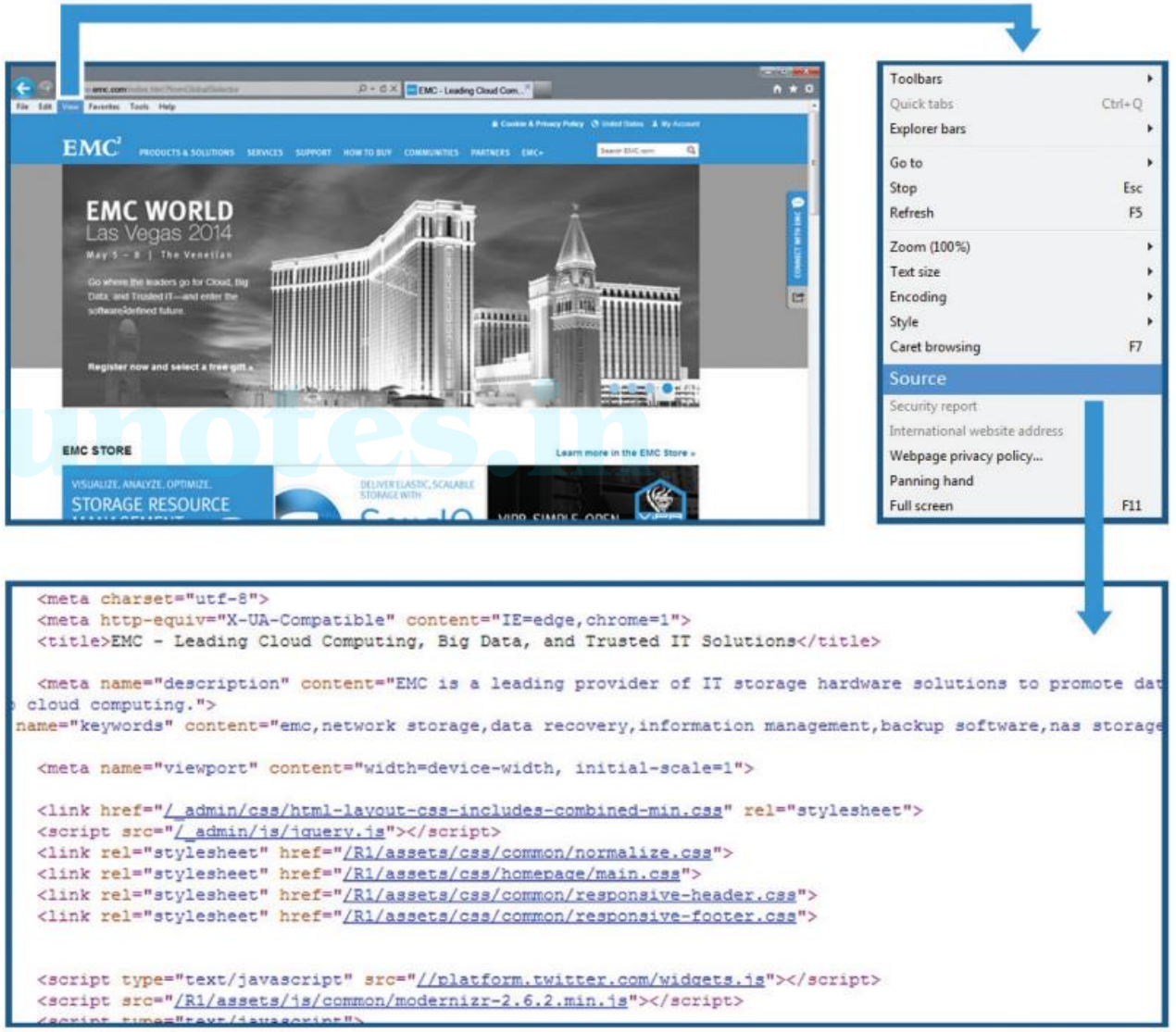
Four main types of data structures – (1/4)

- **Structured data:** Data containing a defined data type, format, and structure (that is, transaction data, online analytical processing [OLAP] data cubes, traditional RDBMS, CSV files, and even simple spread sheets).

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

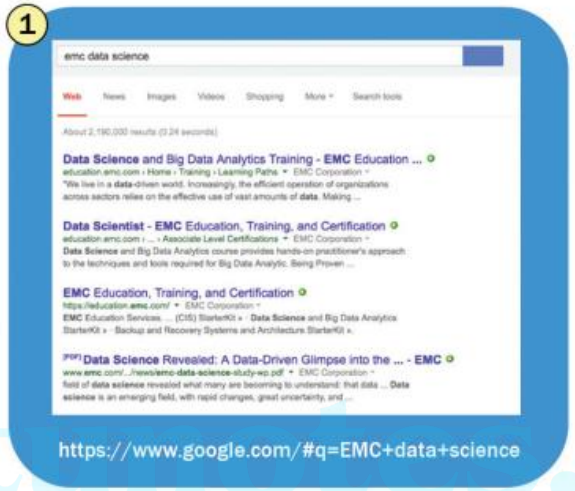
Four main types of data structures - (2/4)

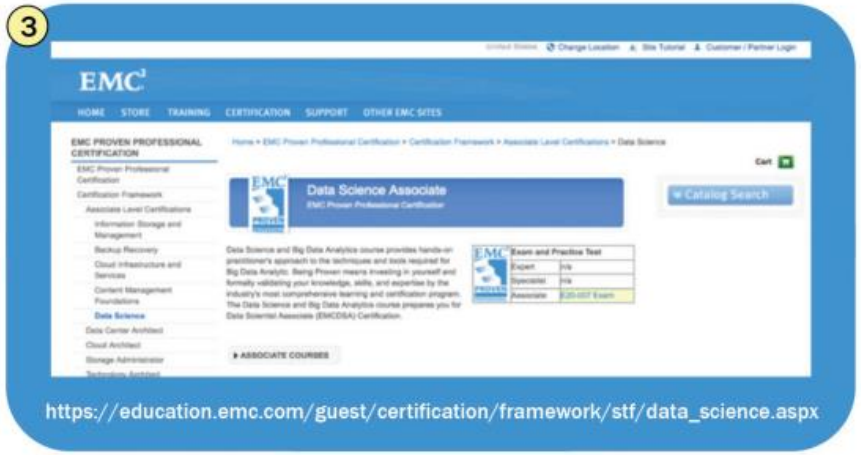
- **Semi-structured data:** Textual data files with a discernible pattern that enables parsing (such as Extensible Markup Language [XML] data files that are self-describing and defined by an XML schema).



Four main types of data structures - (3/4)

- **Quasi-structured data:** Textual data with erratic data formats that can be formatted with effort, tools, and time (for instance, web clickstream data that may contain inconsistencies in data values and formats).





Four main types of data structures - (4/4)

- **Unstructured data:** Data that has no inherent structure, which may include text documents, PDFs, images, and video.

Analyst Perspective on Data Repositories

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

State of the Practice in Analytics

State of the Practice in Analytics

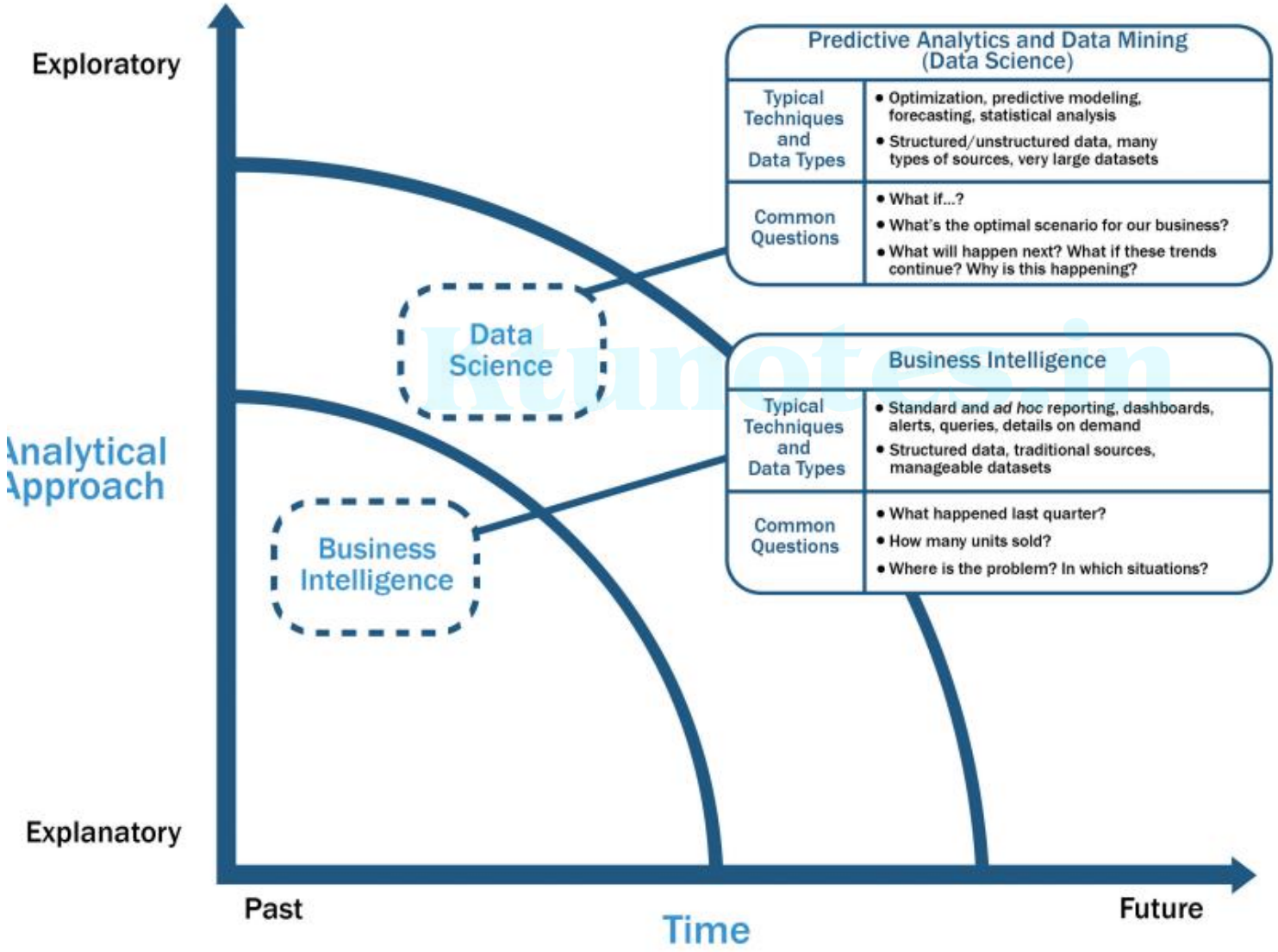
- Current business problems provide many opportunities for organizations to become more analytical and data driven

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)

State of the Practice in Analytics

- Organizations can apply advanced analytical techniques to optimize processes and derive more value from these common tasks.
- The first three examples do not represent new problems.
  - Organizations have been trying to reduce customer churn(movement), increase sales, and cross-sell (sell a different product or service to an existing customer) customers for many years.
  - Persuade a customer to buy something additional or more expensive – Upsell
- The last example portrays emerging regulatory requirements.
- Many compliance and regulatory laws have been in existence for decades, but additional requirements are added every year, which represent additional complexity and data requirements for organizations.
  - Laws related to anti-money laundering (AML) and fraud prevention require advanced analytical techniques to comply with and manage properly.

BI Versus Data Science



BI Versus Data Science

- One way to evaluate the type of analysis being performed is to examine the time horizon and the kind of analytical approaches being used.
- BI tends to provide reports, dashboards, and queries on business questions for the current period or in the past.
- BI systems make it easy to answer questions related to
  - quarter-to-date revenue
  - progress toward quarterly targets
  - how much of a given product was sold in a prior quarter or year
- closed-ended and explain current or past behavior
- BI provides hindsight and some insight and generally answers questions related to “when” and “where” events occurred.

BI Versus Data Science

- Data Science
  - use disaggregated data
  - focusing on analyzing the present
  - enabling informed decisions about the future.
  - more exploratory in nature
  - may use scenario optimization to deal with more open-ended questions
  - focusing on questions related to “how” and “why” events occur
- Rather than aggregating historical data to look at how many of a given product sold in the previous quarter, a team may employ Data Science techniques such as time series analysis, to forecast future product sales and revenue more accurately than extending a simple trend line.
- BI problems tend to require highly structured data organized in rows and columns for accurate reporting

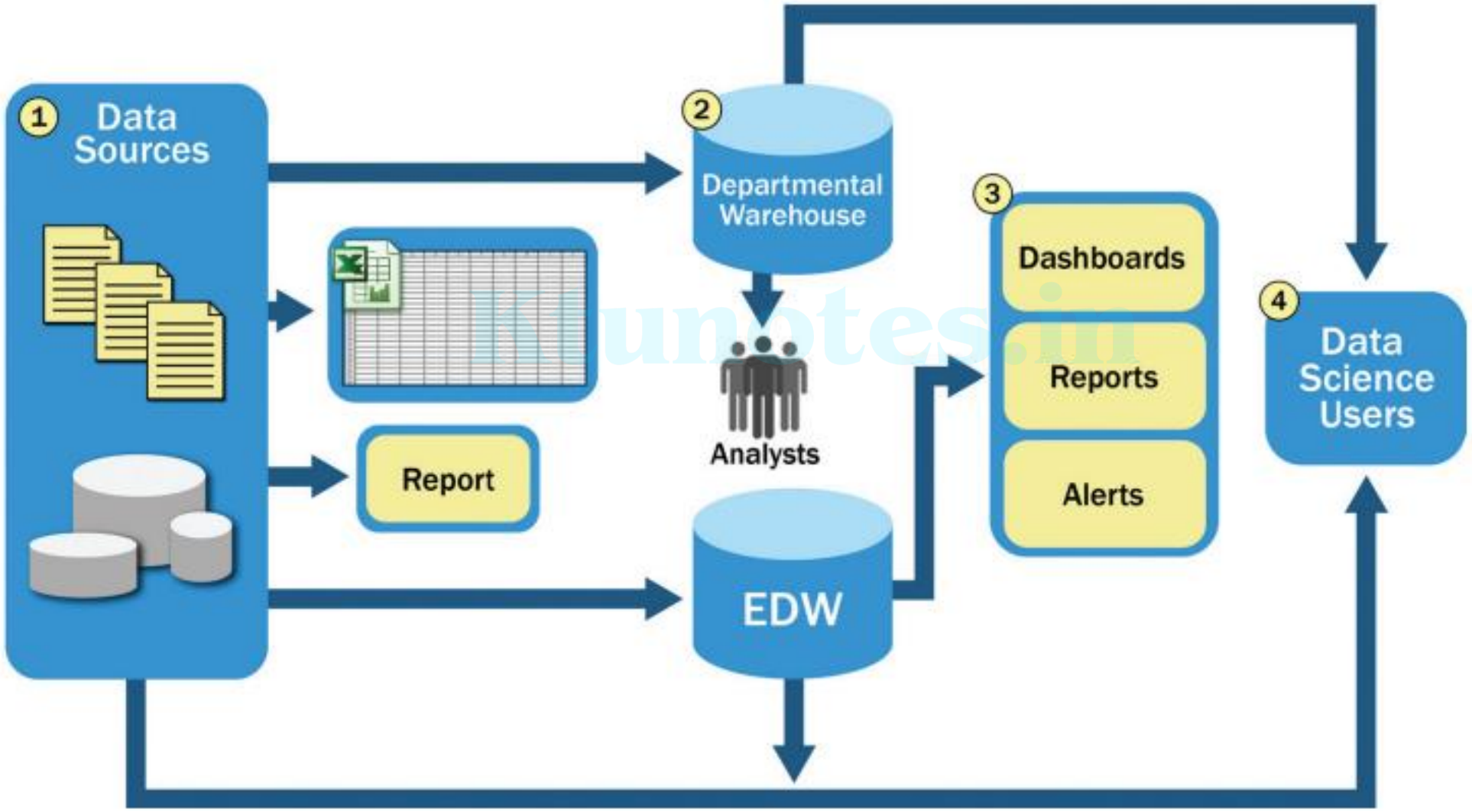
BI Versus Data Science

- Data Science projects tend to use many types of data sources, including large or unconventional datasets.
- Depending on an organization's goals, it may choose to embark on a BI project if it is doing reporting, creating dashboards, or performing simple visualizations, or it may choose Data Science projects if it needs to do a more sophisticated analysis with disaggregated or varied datasets.

Current Analytical Architecture

- Data Science projects need workspaces that are purposebuilt for experimenting with data, with flexible and agile data architectures.
- Most organizations still have data warehouses that provide excellent support for traditional reporting and simple data analysis activities but unfortunately have a more difficult time supporting more robust analyses.
- Data flow to the Data Scientist and how this individual fits into the process of getting data to analyze on projects

Current Analytical Architecture



Current Analytical Architecture

1. For data sources to be loaded into the data warehouse, data needs to be well understood, structured, and normalized with the appropriate data type definitions.

- Although this kind of centralization enables security, backup, and failover of highly critical data,
  - data typically must go through significant preprocessing and checkpoints before it can enter this sort of controlled environment
  - does not lend itself to data exploration and iterative analytics

Current Analytical Architecture

2. As a result of this level of control on the EDW, additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis.
   - These local data marts may not have the same constraints for security and structure as the main EDW and allow users to do some level of more in-depth analysis.
   - However, these one-off systems reside in isolation, often are not synchronized or integrated with other data stores, and may not be backed up.

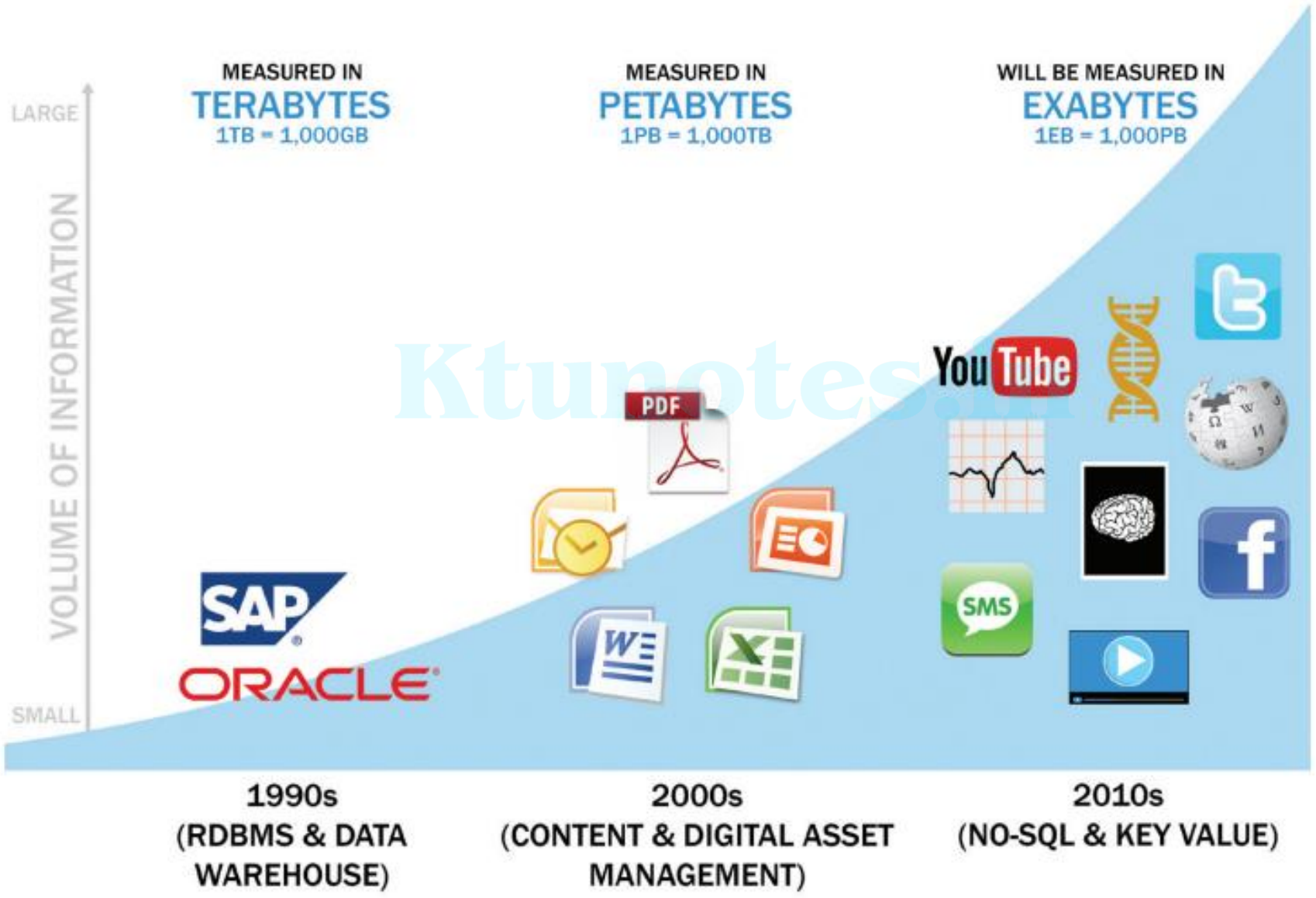
Current Analytical Architecture

3. Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes.
   - These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.

Current Analytical Architecture

4. At the end of this workflow, analysts get data provisioned for their downstream analytics.
   - Because users generally are not allowed to run custom or intensive analytics on production databases, analysts create data extracts from the EDW to analyze data offline in R or other local analytical tools.
   - Many times these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.
   - Because these analyses are based on data extracts, they reside in a separate location, and the results of the analysis—and any insights on the quality of the data or anomalies—rarely are fed back into the main data repository

Drivers of Big Data



Drivers of Big Data

- The data now comes from multiple sources, such as these:
  - Medical information, such as genomic sequencing and diagnostic imaging
  - Photos and video footage uploaded to the World Wide Web
  - Video surveillance, such as the thousands of video cameras spread across a city
  - Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
  - Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
  - Nontraditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

Emerging Big Data Ecosystem & a New Approach to Analytics

- Organizations and data collectors are realizing that the data they can gather from individuals contains intrinsic value and, as a result, a new economy is emerging.
- As this new digital economy continues to evolve, the market sees the introduction of data vendors and data cleaners that use crowdsourcing (such as Mechanical Turk and GalaxyZoo) to test the outcomes of machine learning techniques.
- Other vendors offer added value by repackaging open source tools in a simpler way and bringing the tools to market.
- Vendors such as Cloudera, Hortonworks, and Pivotal have provided this value-add for the open source framework Hadoop.

Emerging Big Data Ecosystem & a New Approach to Analytics

- As the new ecosystem takes shape, there are four main groups of players within this interconnected web.
  - Data Devices
  - Data collectors
  - Data Aggregators
  - Data users and buyers

Data devices

- Gather data from multiple locations and continuously generate new data about this data.
- For each gigabyte of new data created, an additional petabyte of data is created about that data.
- Online video game through a PC, game console, or smartphone, Retail shopping loyalty cards record

Data collectors

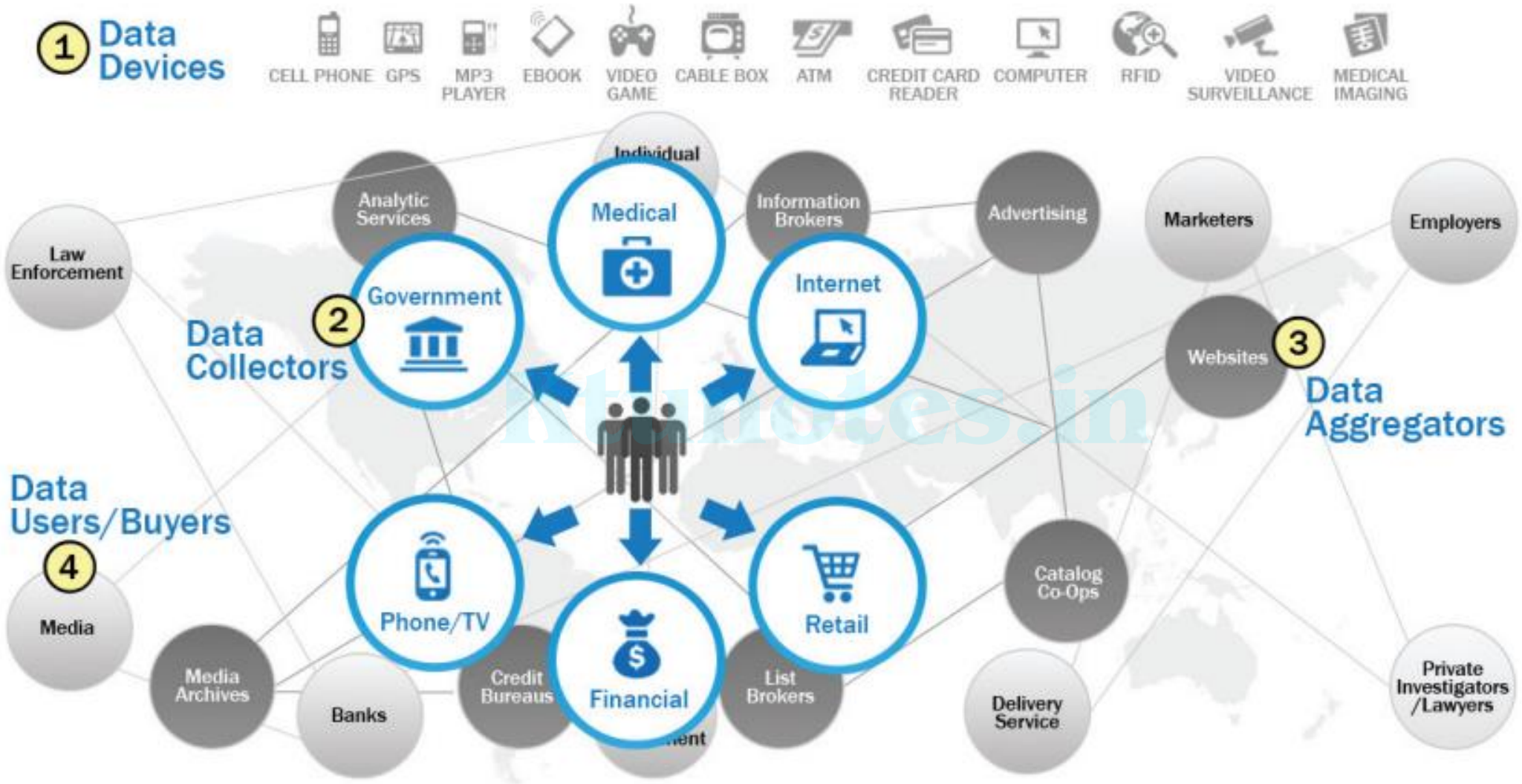
- Include sample entities that collect data from the device and users
- Data results from a cable TV provider tracking the shows a person watches, which TV channels someone will and will not pay for to watch on demand, and the prices someone is willing to pay for premium TV content
- Retail stores tracking the path a customer takes through their store while pushing a shopping cart with an RFID chip so they can determine which products get the most foot traffic using geospatial data collected from the RFID chips

Data aggregators

- make sense of the data collected from the various entities from the "SensorNet" or the "Internet of Things."
- These organizations compile data from the devices and usage patterns collected by government agencies, retail stores, and websites.
- In turn, they can choose to transform and package the data as products to sell to list brokers, who may want to generate marketing lists of people who may be good targets for specific ad campaigns.

Data users and buyers

- These groups directly benefit from the data collected and aggregated by others within the data value chain.
- Retail banks, acting as a data buyer, may want to know which customers have the highest likelihood to apply for a second mortgage or a home equity line of credit.
- This kind of data may include demographic information about people living in specific locations; people who appear to have a specific level of debt, yet still have solid credit scores (or other characteristics such as paying bills on time and having savings accounts) that can be used to infer credit worthiness; and those who are searching the web for information about paying off debts or doing home remodeling projects.

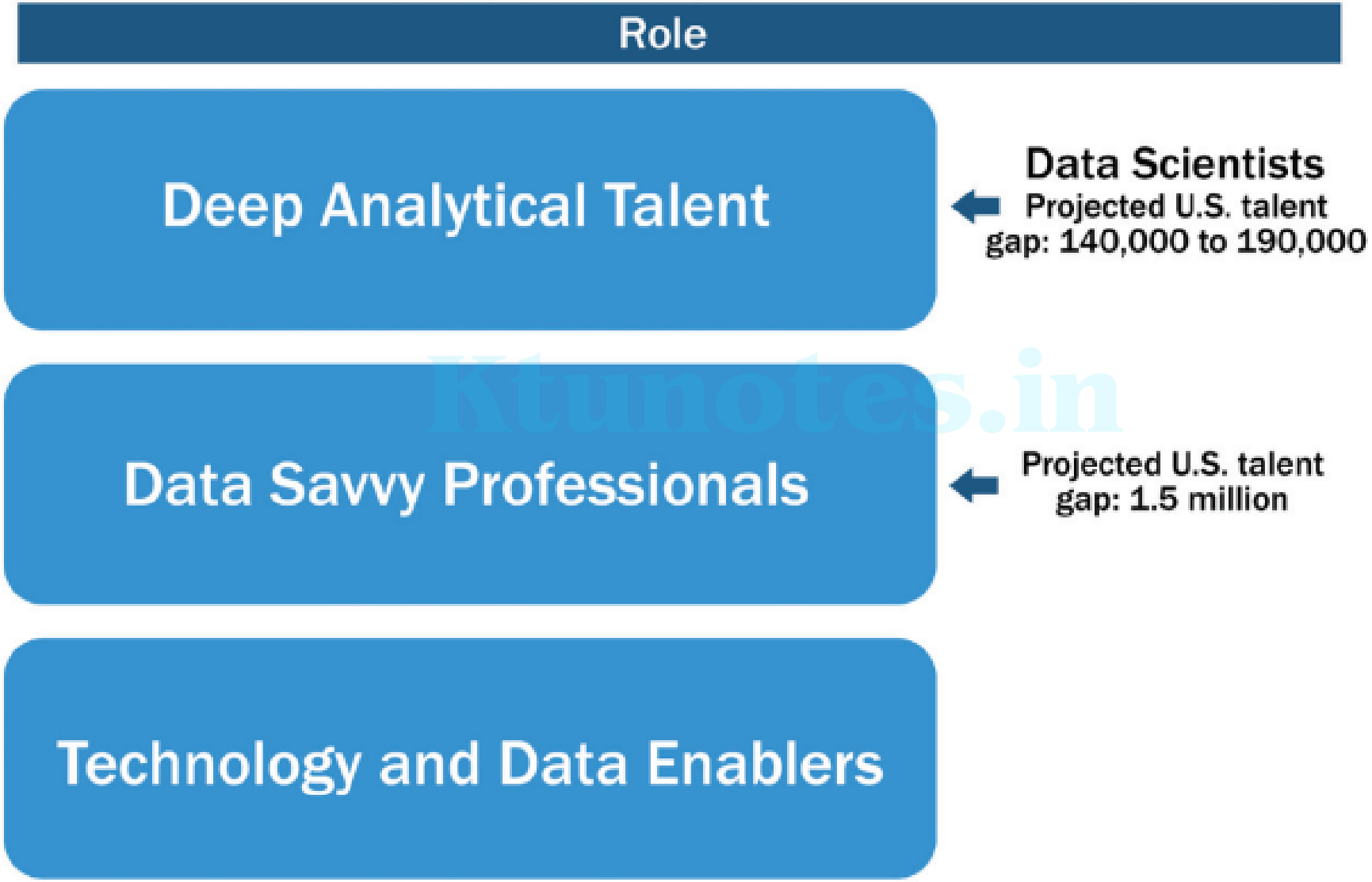


Emerging Big Data ecosystem

Key Roles for the New Big Data Ecosystem

- New players have emerged to curate, store, produce, clean, and transact data.
- Need for applying more advanced analytical techniques to increasingly complex business problems has driven the emergence of new roles, new technology platforms, and new analytical methods.

Three Key Roles of The New Data Ecosystem



Deep Analytical Talent

- Technically savvy, with strong analytical skills.
- Members possess a combination of skills to handle raw, unstructured data and to apply complex analytical techniques at massive scales.
- This group has advanced training in quantitative disciplines, such as mathematics, statistics, and machine learning.
- Need access to a robust analytic sandbox or workspace where they can perform large-scale analytical data experiments.
- Example: statisticians, economists, mathematicians, and the new role of the Data Scientist.

Data Savvy Professionals

- less technical depth but has a basic knowledge of statistics or machine learning and can define key questions that can be answered using advanced analytics.
- These people tend to have a base knowledge of working with data, or an appreciation for some of the work being performed by data scientists and others with deep analytical talent.
- Examples of data savvy professionals include financial analysts, market research analysts, life scientists, operations managers, and business and functional managers.
- At a high level, this means for every Data Scientist profile needed, the gap will be ten times as large for Data Savvy Professionals.

Technology and Data Enablers

- People providing technical expertise to support analytical projects, such as provisioning and administrating analytical sandboxes, and managing large-scale data architectures that enable widespread analytics within companies and other organizations.
- requires skills related to computer engineering, programming, and database administration.
- These three groups must work together closely to solve complex Big Data challenges.
- Most organizations are familiar with people in the latter two groups mentioned, but the first group, Deep Analytical Talent, tends to be the newest role for most and the least understood.

Activities of Data Scientist

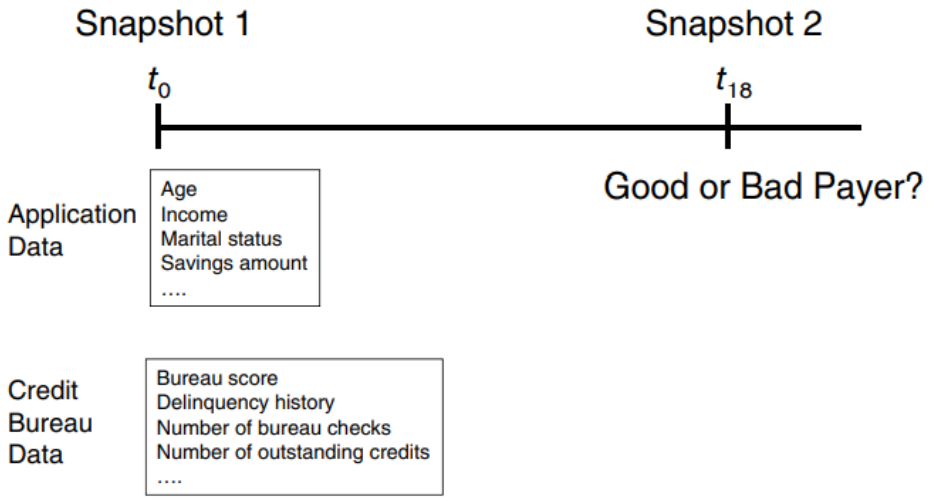
Three recurring sets of activities that data scientists perform:

1. **Reframe business challenges as analytics challenges**
   - Skill to diagnose business problems, consider the core of a given problem, and determine which kinds of candidate analytical methods can be applied to solve it.
2. **Design, implement, and deploy statistical models and data mining techniques on Big Data.**
   - What people think about when they consider the role of the Data Scientist: namely, applying complex or advanced analytical methods to a variety of business problems using data.
3. **Develop insights that lead to actionable recommendations.**
   - It is critical to note that applying advanced methods to data problems does not necessarily drive new business value.
   - Instead, it is important to learn how to draw insights out of the data and communicate them effectively.

Example Applications

CREDIT RISK MODELING

- Different types of analytical models will be built in a credit risk setting.
- A first example are application scorecards.
- These are models that score credit applications based on their creditworthiness.
- They are typically constructed by taking two snapshots of information: application and credit bureau information at loan origination and default status information 12 or 18 months ahead.



CREDIT RISK MODELING

- Logistic regression is a very popular application scorecard construction technique due to its simplicity and good performance.

Characteristic Name	Attribute	Points
Age 1	Up to 26	100
Age 2	26–35	120
Age 3	35–37	185
Age 4	37+	225
Employment status 1	Employed	90
Employment status 2	Unemployed	180
Salary 1	Up to 500	120
Salary 2	501–1,000	140
Salary 3	1,001–1,500	160
Salary 4	1,501–2,000	200
Salary 5	2,001+	240

the following logistic regression with WOE coding was used:

$$P(Customer = good \mid age,\ employment,\ salary) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 WOE_{age} + \beta_2 WOE_{employment} + \beta_3 WOE_{salary})}}$$

Typically, the model will then be re-expressed in terms of the log odds, as follows:

$$\log\left(\frac{P(Customer = good|age,\ employment,\ salary)}{P(Customer = bad|age,\ employment,\ salary)}\right)$$
$$= \beta_0 + \beta_1 WOE_{age} + \beta_2 WOE_{employment} + \beta_3 WOE_{salary}$$

One then commonly applies a scorecard scaling by calculating a score as a linear function of the log odds, as follows:

$$\text{Score} = \text{offset} + \text{factor} * \log(\text{odds})$$

Assume that we want a score of 600 for odds of 50:1, and a score of 620 for odds of 100:1. This gives the following:

$$600 = \text{offset} + \text{factor} * \log(50)$$

$$620 = \text{offset} + \text{factor} * \log(100)$$

The offset and factor then become:

$$\text{factor} = 20/\ln(2)$$

$$\text{offset} = 600 - \text{factor} * \ln(50)$$

Once these values are known, the score becomes:

$$\text{Score} = \left(\sum_{i=1}^{N} (WOE_i * \beta_i) + \beta_0 \right) * \text{factor} + \text{offset}$$

$$\text{Score} = \left(\sum_{i=1}^{N} \left(WOE_i * \beta_i + \frac{\beta_0}{N} \right) \right) * \text{factor} + \text{offset}$$

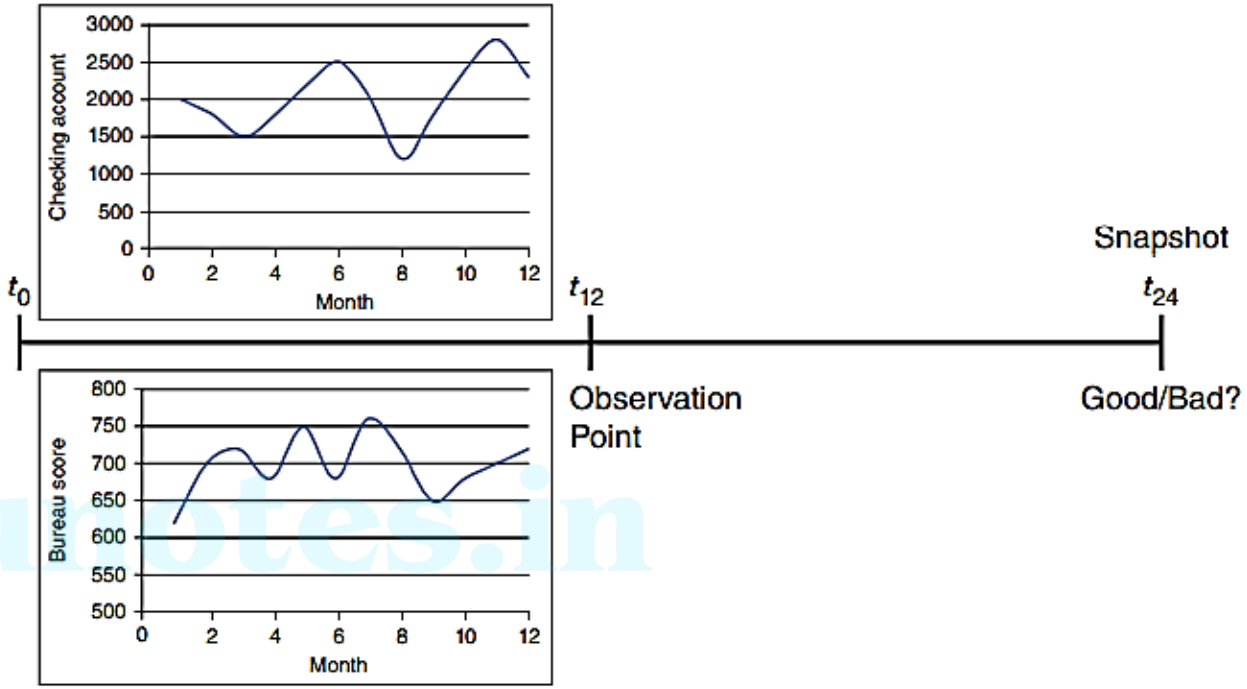
$$\text{Score} = \left(\sum_{i=1}^{N} \left(WOE_i * \beta_i + \frac{\beta_0}{N} \right) * \text{factor} + \frac{\text{offset}}{N} \right)$$

Hence, the points for each attribute are calculated by multiplying the weight of evidence of the attribute with the regression coefficient of the characteristic, then adding a fraction of the regression intercept, multiplying the result by the factor, and finally adding a fraction of the offset

Behavioral scorecards

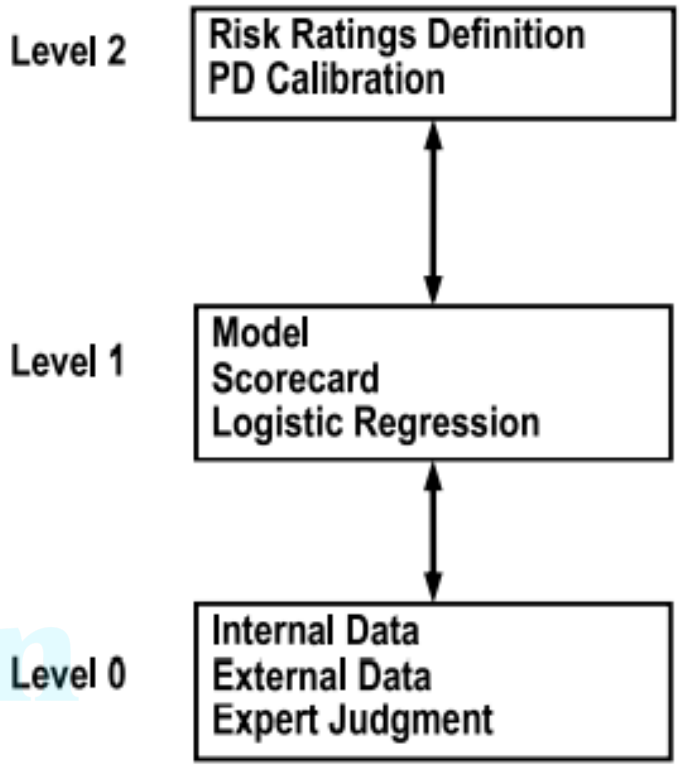
- These are analytical models that are used to score the default behavior of an existing portfolio of customers.
- On top of the application characteristics, behavioral characteristics, such as trends in account balance or bureau score, delinquency history, credit limit increase/decrease, and address changes, can also be used.
- Because behavioral scorecards have more data available than application scorecards, their performance (e.g., measured using AUC) will be higher.
- Next to debt provisioning, behavioral scorecards can also be used for marketing (e.g., up/down/cross-selling) and/or proactive debt collection.

- Both application and behavioral scorecards are then used to calculate the probability of default (PD) for a portfolio of customers.
- This is done by first segmenting the scores into risk ratings and then calculating a historically observed default rate for each rating, which is then used to project the probability of default (PD) for (typically) the upcoming year.



Constructing a Data Set for Behavioral Scoring

Three Level Credit Risk Model



- Other measures that need to be calculated in credit risk modeling are the loss given default (LGD) and exposure at default (EAD).
- LGD measures the economic loss expressed as a percentage of the outstanding loan amount and is typically estimated using linear regression or regression trees.
- EAD represents the outstanding balance for on-balance sheet items (e.g., mortgages, installment loans).

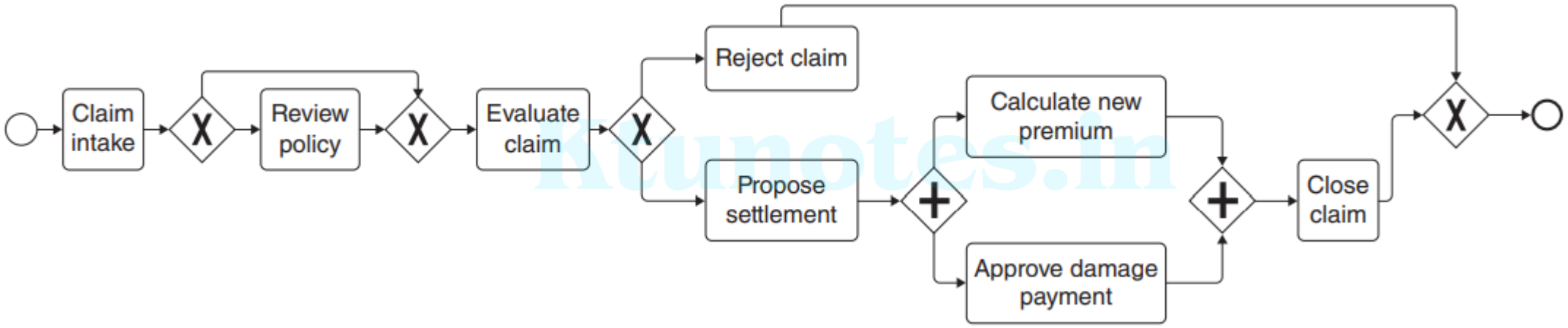
- For off- balance sheet items (e.g., credit cards, credit lines), the EAD is typically calculated as follows: EAD = DRAWN + CCF * (LIMIT−DRAWN), whereby DRAWN represents the already drawn balance, LIMIT the credit limit, and CCF the credit conversion factor, which is expressed as a percentage between 0 and 1.

BUSINESS PROCESS ANALYTICS

BUSINESS PROCESS ANALYTICS

- The management field aims to provide an encompassing approach in order to align an organization's business processes with the concerns of every involved stakeholder
- A business process is then a collection of structured, interrelated activities or tasks that are to be executed to reach a particular goal (produce a product or deliver a service).
- Involved parties in business processes include, among others, managers ("process owners"), who expect work to be delegated swiftly and in an optimal manner; employees, who desire clear and understandable guidelines and tasks that are in line with their skillset; and clients who, naturally, expect effi ciency and quality results from their suppliers.

Example business process model for an insurance claim intake process shown in the business process modeling language (BPMN) standard

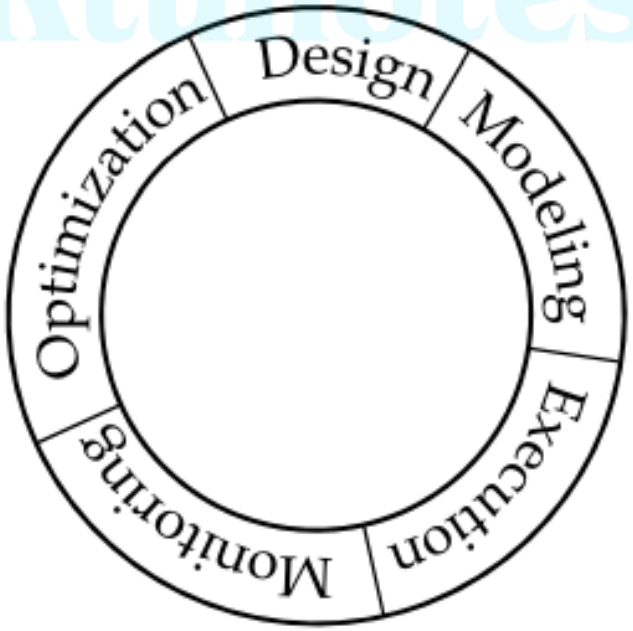


Business Process Management(BPM)

- BPM is oftentimes described as a “process optimization” methodology and is therefore mentioned together with related quality control terms such as total quality management (TQM), six sigma efforts, or continuous process improvement methodologies.
- One significant focal point of BPM is the actual improvement and optimization of processes, but the concept also encompasses best practices toward the design and modeling of business processes, monitoring (consider for instance compliance requirements), and gaining insights by unleashing analytical tools on recorded business activities.

Business process lifecycle

- All these activities are grouped within the “business process lifecycle,” starting with the design and analysis of a business process (modeling and validation), its configuration (implementation and testing), its enactment (execution and monitoring), and finally, the evaluation, which in turn leads again to the design of new processes



Process Intelligence

- It is mainly in the last part of the BPM life cycle (i.e., evaluation) where the concepts of process analytics and process intelligence fit in.
- process intelligence is a very broad term describing a plethora of tools and techniques, and can include anything that provides information to support decision making
- Another term that has become commonplace in a process intelligence context is business activity monitoring (BAM), which refers to real-time monitoring of business processes and immediate reaction if a process displays a particular pattern.
- Corporate performance management (CPM) is another popular term for measuring the performance of a process or the orga□nization as a whole.

Process Mining and Analytics

- a new research fi eld has emerged, denoted as “process mining,” which positions itself between BPM and traditional data mining.
- The discipline aims to provide a comprehensive set of tools to provide process-centered insights and to drive process
- process mining builds on existing approaches, such as data mining and model-driven approaches, but is more than just the sum of these components
- traditional existing data mining techniques are too data-centric to provide a solid understanding of the end-to-end processes in an organization, whereas business intelligence tools focus on simple dashboards and reporting. It is exactly this gap that is narrowed by process mining tools, thus enabling true business process analytics.

Process Mining and Analytics

- The most common task in the area of process mining is called process discovery, in which analysts aim to derive an as-is process model starting from the data as it is recorded in process-aware information support systems instead of starting from a to-be descriptive model, and trying to align the actual data to this model.
- A significant advantage of process discovery is the fact that only a limited amount of initial data is required to perform a fi rst exploratory analysis.

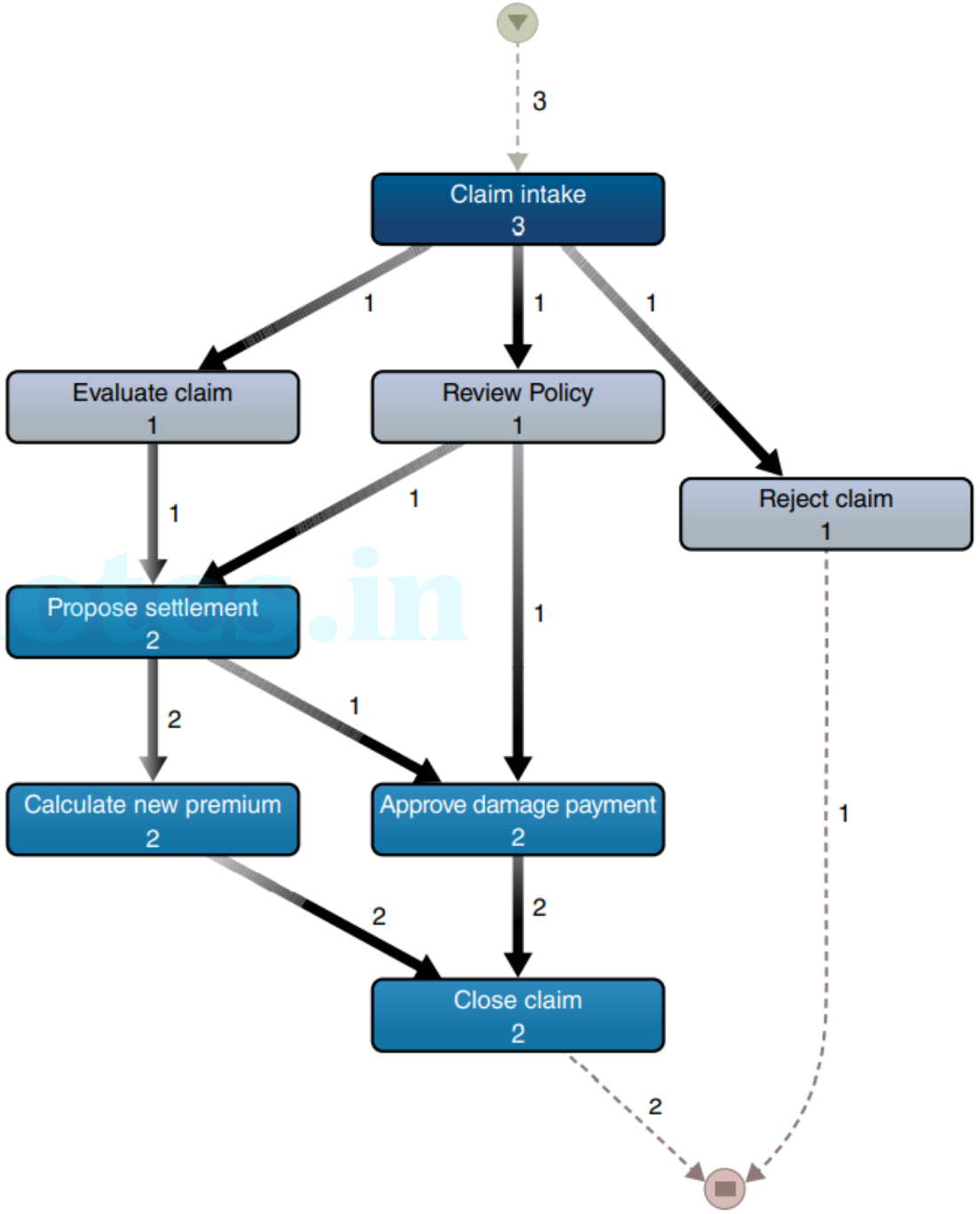
Process Mining and Analytics

- To perform a process discovery task, we start our analysis from a so-called "event log": a data table listing the activities that have been executed during a certain time period, together with the case (the process instance) to which they belong.
- Next, an algorithm iterates over all process cases and creates "flows of work" between the activities.

Example Insurance Claim Handling Event Log

Case Identifier	Start Time	Completion Time	Activity
Z1001	8-13-2013 09:43:33	8-13-2013 10:11:21	Claim intake
Z1004	8-13-2013 11:55:12	8-13-2013 15:43:41	Claim intake
Z1001	8-13-2013 14:31:05	8-16-2013 10:55:13	Evaluate claim
Z1004	8-13-2013 16:11:14	8-16-2013 10:51:24	Review policy
Z1001	8-17-2013 11:08:51	8-17-2013 17:11:53	Propose settlement
Z1001	8-18-2013 14:23:31	8-21-2013 09:13:41	Calculate new premium
Z1004	8-19-2013 09:05:01	8-21-2013 14:42:11	Propose settlement
Z1001	8-19-2013 12:13:25	8-22-2013 11:18:26	Approve damage payment
Z1004	8-21-2013 11:15:43	8-25-2013 13:30:08	Approve damage payment
Z1001	8-24-2013 10:06:08	8-24-2013 12:12:18	Close claim
Z1004	8-24-2013 12:15:12	8-25-2013 10:36:42	Calculate new premium
Z1011	8-25-2013 17:12:02	8-26-2013 14:43:32	Claim intake
Z1004	8-28-2013 12:43:41	8-28-2013 13:13:11	Close claim
Z1011	8-26-2013 15:11:05	8-26-2013 15:26:55	Reject claim

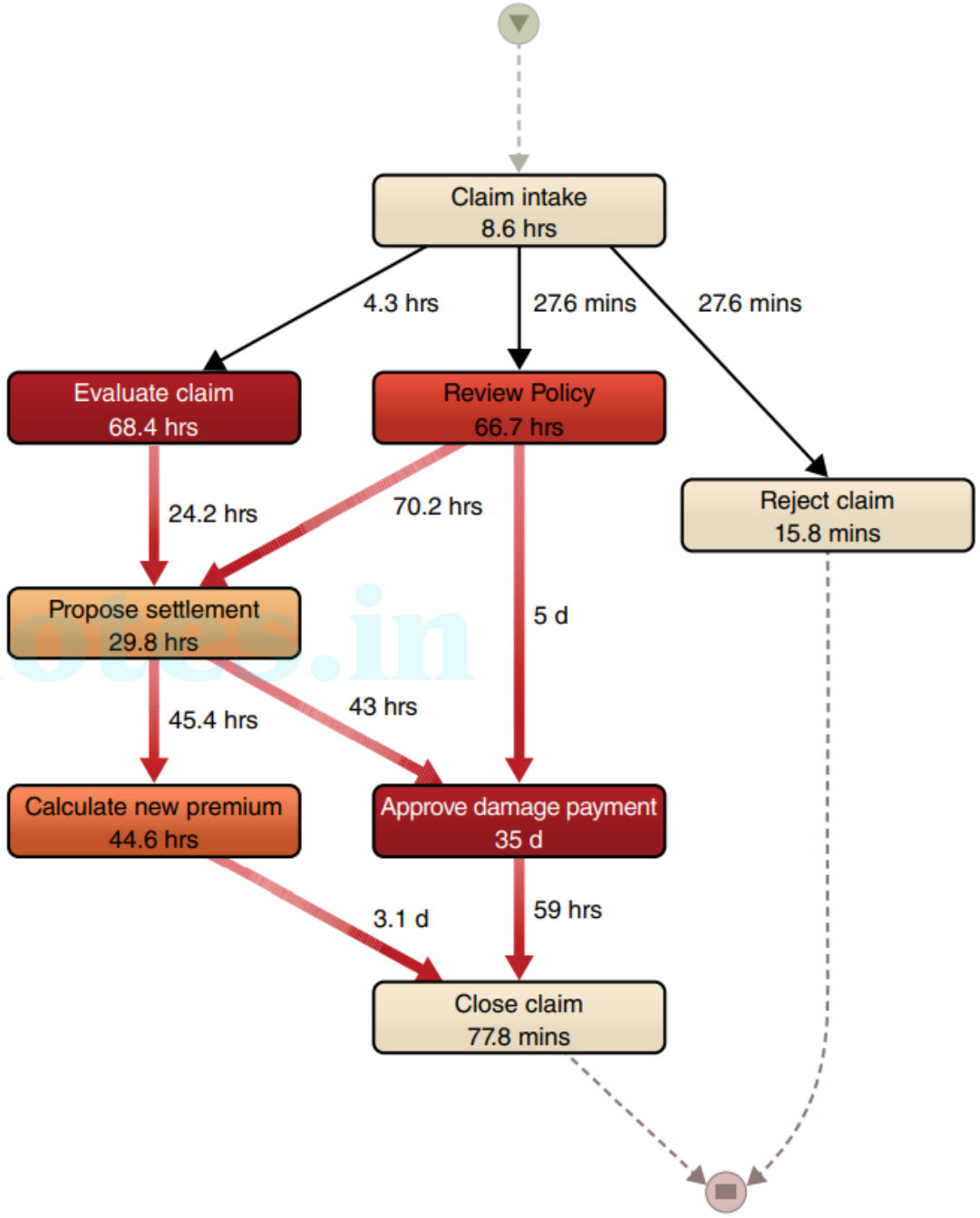
- After executing the process discovery algorithm, a process map can be obtained
- Activities that follow each other distinctly (no overlapping start and end times) will be put in a sequence.
- When the same activity is followed by different activities over various process instances, a split is created
- When two or more activities' executions overlap in time, they are executed in parallel and are thus both fl owing from a common predecessor.



- The process map can be annotated with various information, such as frequency counts of an activity's execution

- Process discovery provides an excellent means to perform an initial exploratory analysis of the data at hand, showing actual and true information.

- This allows practitioners to quickly determine bottlenecks, deviations, and exceptions in the day-to-day workfl ows.



Example Process Map Annotated with Performance Information

Thank you...

Compiled by,
Mr. Shyam Krishna K
Asst. Professor, Dept. of CSE
Sahrdaya CET
shyamkrishna@sahrdaya.ac.in