Big Data For Analysis Overview Training

15-17 October 2014

Organized By:

MDEC
Driving Transformation

Partners:

InfoTrek

aBy res
Absolute Binary Resources

About: Me

Mohd Izhar Firdaus Ismail

- Solution Architect of ABYRES Sdn Bhd, and Hortonworks Certified Instructor
- 7 years involvement in helping people realize value on Open Source software through Fedora Community as Fedora Ambassador (Volunteer)
- 5 years experience in software development, infrastructure management, and training for knowledge management, and resource sharing platforms for United Nations ILO, Center of Internet and Society India, World Council of Churches
- 7 years experience in utilizing Python programming language for both system development and data analysis

Before we start …

What is your expectations coming here?
What do you wish to get by the end of this training?

Introduce yourself

- Name
- Ministry / Agency
- Role / Position in Ministry / Agency
- What is your expectations coming here?
- What do you wish to get by the end of this training?

Lets make this an engaging session

Feel free to ask questions and start discussions

Day 1:
Introduction to
Big Data, Hadoop and Data Science

Day 1 : Objectives

- Understand what is Data Science and the types of expertise needed in it
- Understand a general data analysis workflow and the common tasks involved in it
- Understand the general history of data analysis, business intelligence, and their enabling technologies.
- Understand the evolution of technologies that lead to Big Data
- Understand what is Big Data Analytics, and what challenges can be considered Big Data and what are not
- Understand what is Hadoop, what it is for in a Big Data infrastructure, when you need it, and when you don't need it
- Know what are available technologies in a Big Data ecosystem

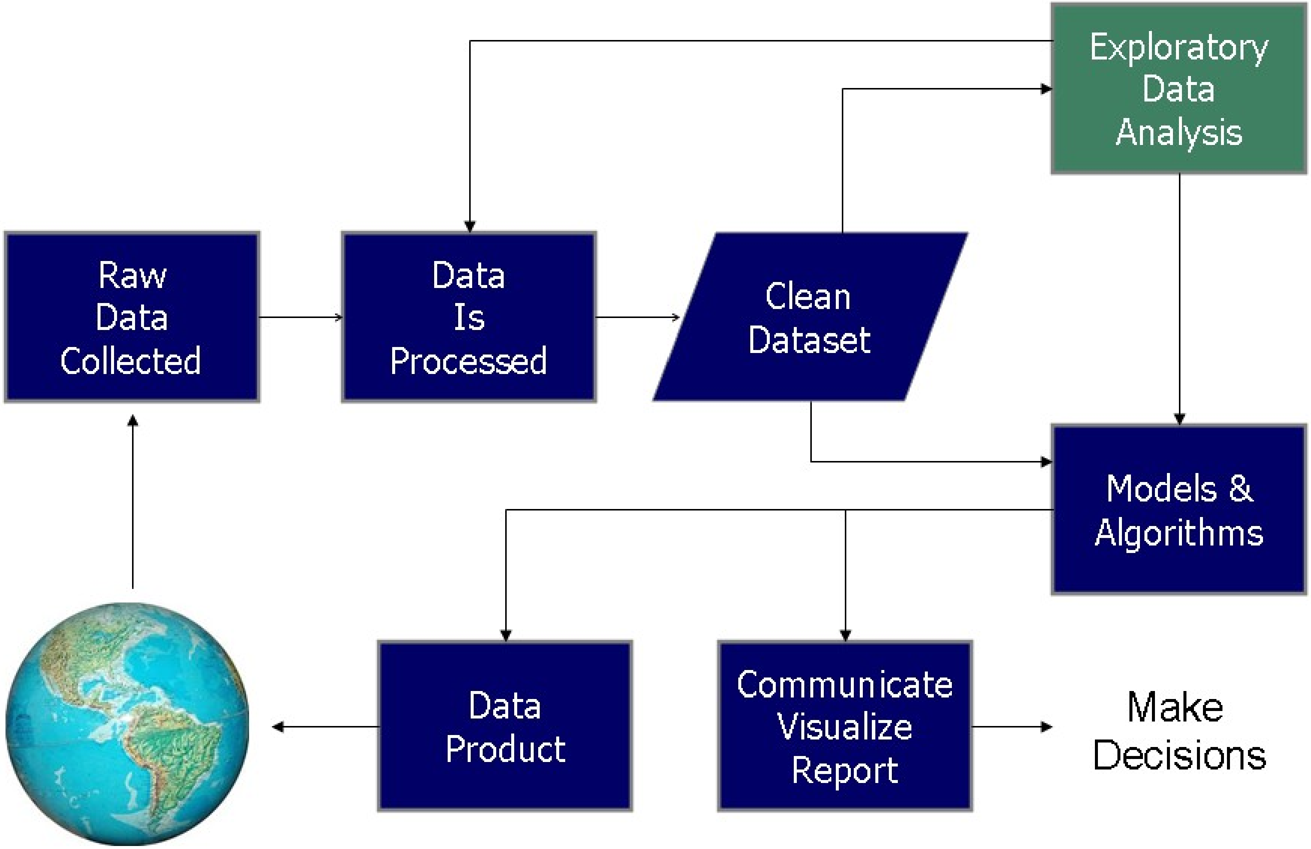
Introduction to Data Analysis

Definition

- Data Analysis is the process of:
    - Inspecting data
    - Cleaning data
    - Transforming data
    - Modeling data

- Data Analysis goal is to:
    - Discover useful information
    - Suggesting conclusion
    - Supporting decision making

Data Science Process



Data Transformation

- Transformation is the core of data analysis
  - Data cleaning & treatment
  - Data modeling
  - Statistical transformations
  - Analysis

Data Cleaning

- The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors
- Common tasks includes:
  - Record matching
  - Deduplication
  - Column segmentation

Data Modeling

- Data modeling is a process used to define and analyze data requirements needed to support the business processes within the scope of corresponding information systems in organizations.
- Activities includes but not limited to:
  - Segregating data into business/functional lines
  - Integrating data into more unified datasets
  - Designing data models to aid with analytical activities

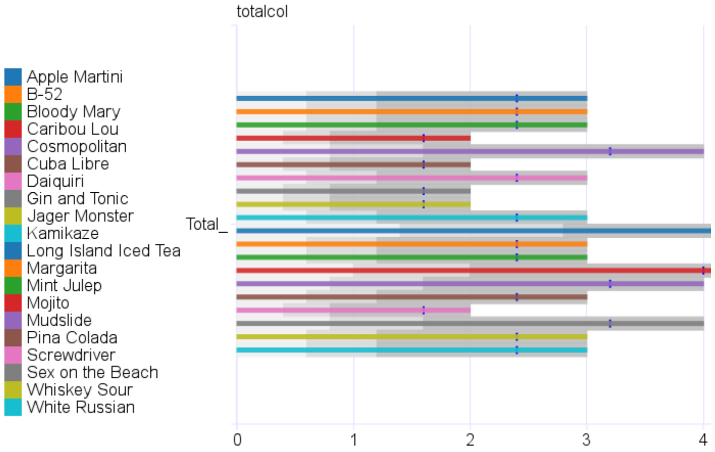
Data Visualization

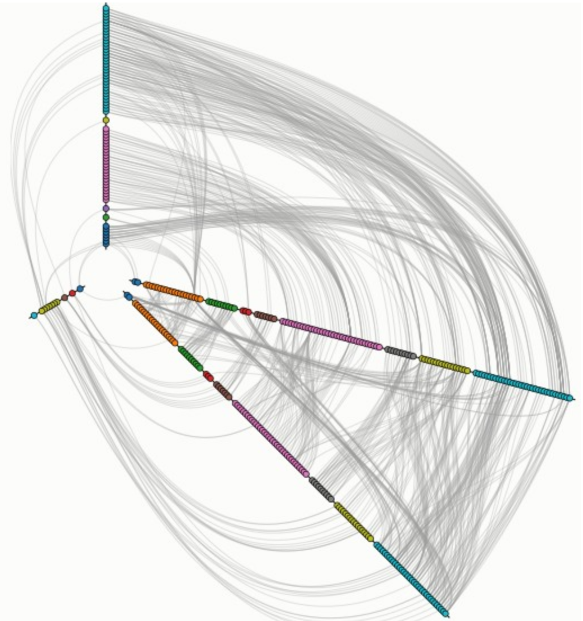
- Generates visual reports and dashboards for data exploration to gain insights
- Comprises of traditional visualization such as charts, maps and graphs, to more complex visualization such as tree diagrams , dendogram, voronoi diagrams, etc

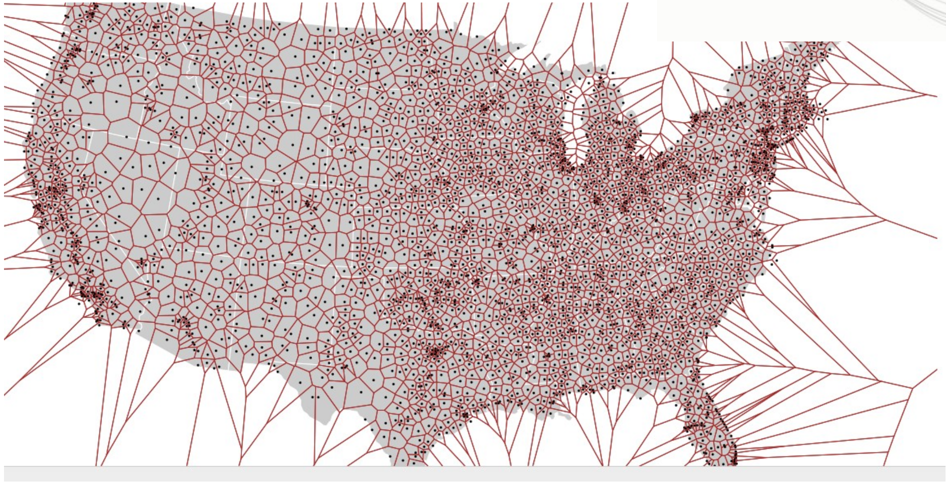
Data Product

- A data product is a product that facilitates an end goal through the use of data.
- Examples:
  - Enabling technologies for data analysis
    - Artificial intelligence software
    - Data analysis products with pre-made data analysis capabilities
  - End user products
    - Google Search, Waze, Google Analytics, etc

flare
vis
analytics
animate
data
display
flex
physics
query
scale
util
axis
controls
data
events
legend
operator
cluster
graph
optimization
FlareVis

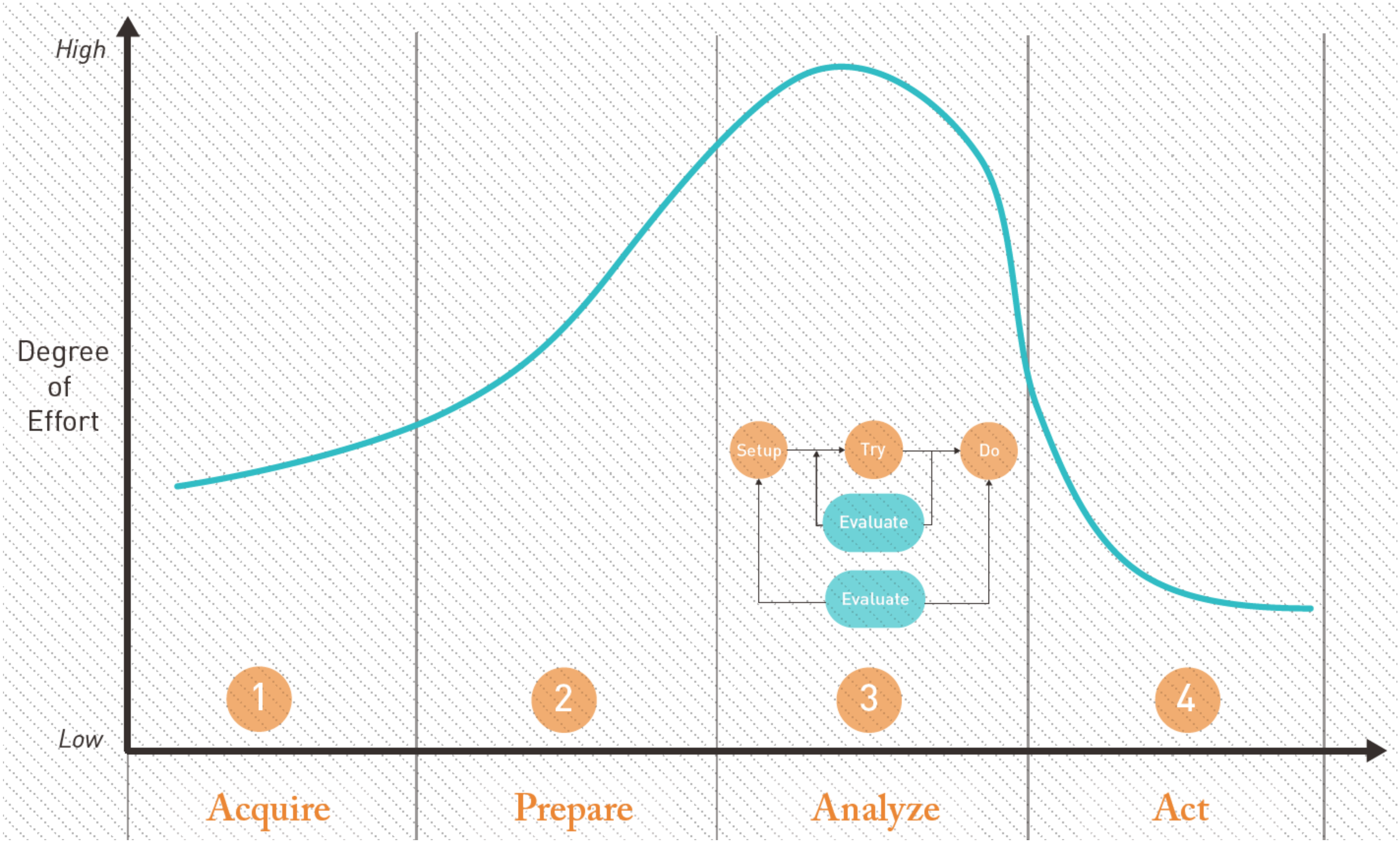

totalcol
Total_
0
1
2
3
4
🟦 Apple Martini
🟧 B-52
🟩 Bloody Mary
🟥 Caribou Lou
🟪 Cosmopolitan
🟫 Cuba Libre
Daiquiri
Gin and Tonic
Jager Monster
Kamikaze
🟦 Long Island Iced Tea
🟧 Margarita
🟩 Mint Julep
🟥 Mojito
🟪 Mudslide
🟫 Pina Colada
Screwdriver
Sex on the Beach
Whiskey Sour
White Russian





Data Science

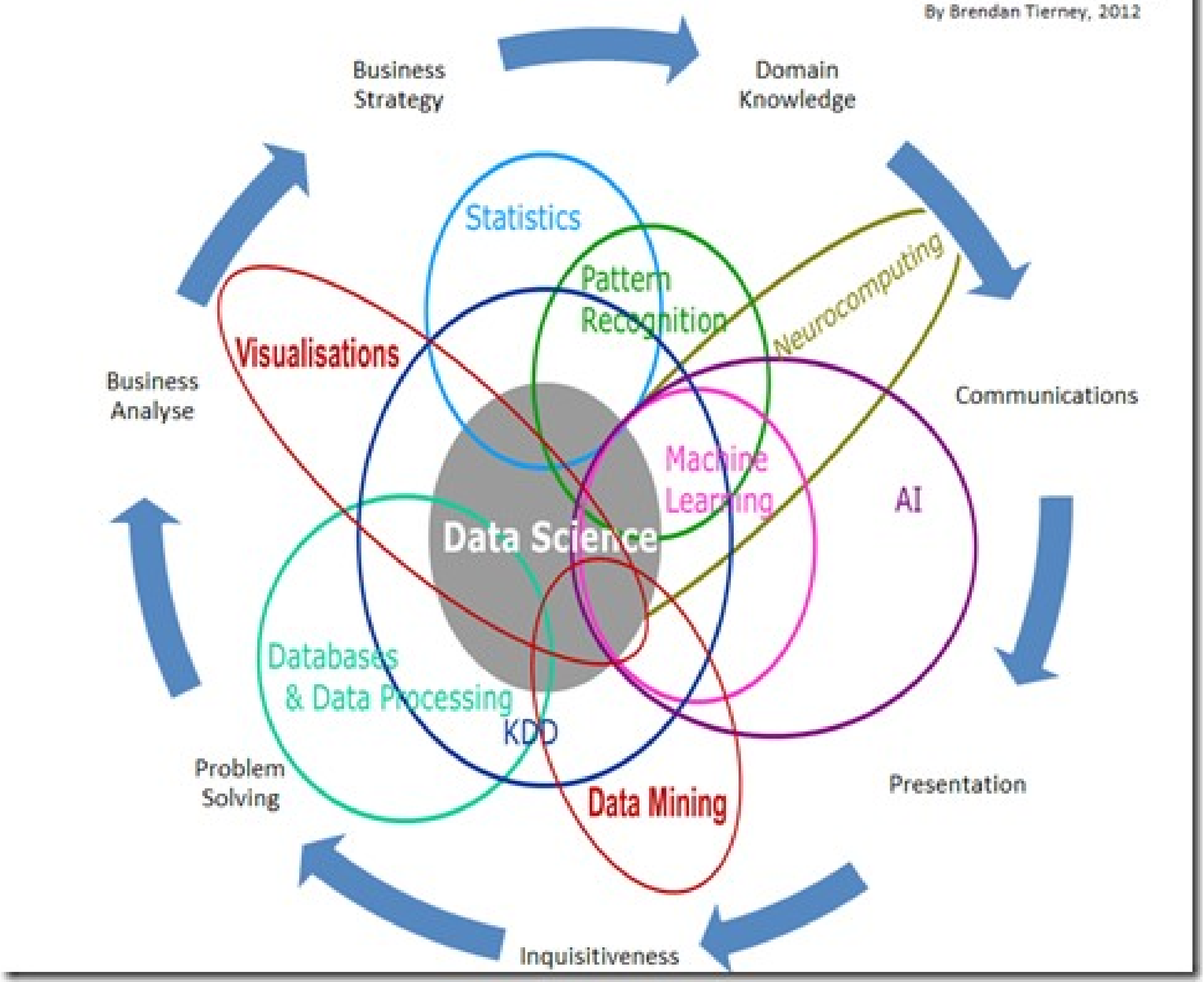
Data science is the study of the generalizable extraction of knowledge from data


High
Degree of Effort
Low
Setup
Try
Do
Evaluate
Evaluate
1
2
3
4
Acquire
Prepare
Analyze
Act

Scientific Process

- Hypothesis
- Test hypothesis
  - Gather data
  - Analyze data
  - Visualize data and create reports
- Derive inference and findings from analyzed data
- Repeat (continuous improvements)

Data Science Is Multidisciplinary



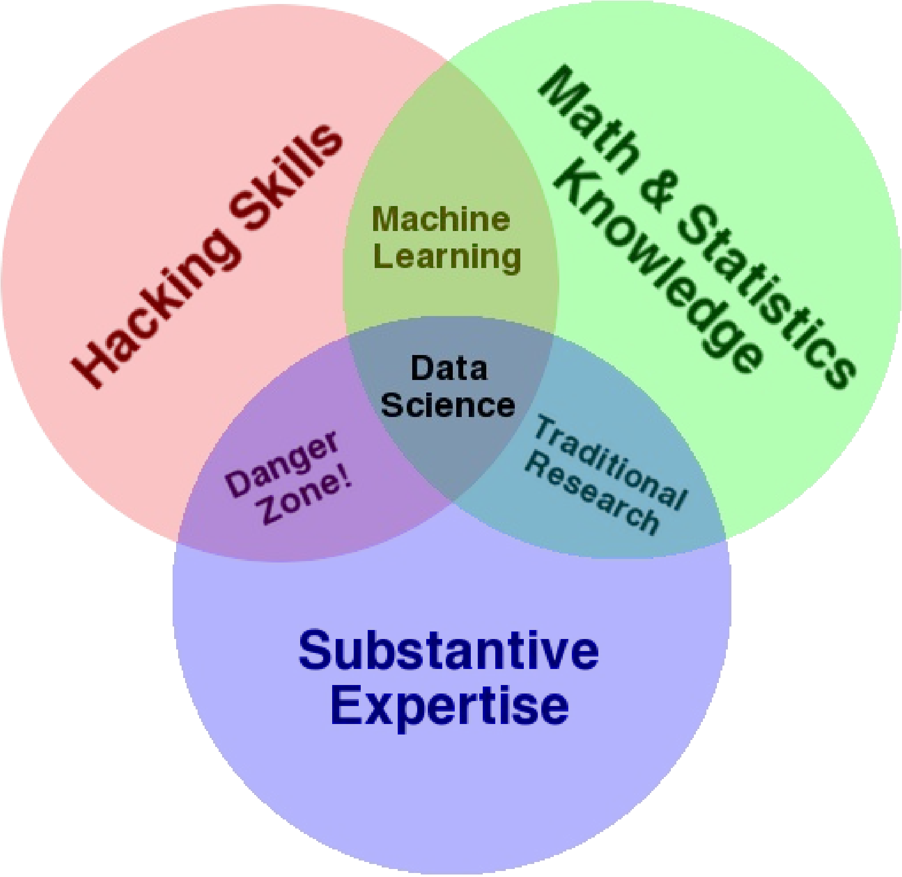
Data Science Disciplines

- Signal processing
- Mathematics
- Probability models
- Machine learning
- Statistical learning
- Computer programming
- Data engineering
- Pattern recognition and learning
- Visualization
- Uncertainty modeling
- Data warehousing
- High performance computing

Goals of Data Science Activity

- Extracting knowledge and information from data
- Creating data products
  - Applications
  - Solutions
  - Devices
  - etc

Data Scientist?



Role and Responsibilities of Data Scientist

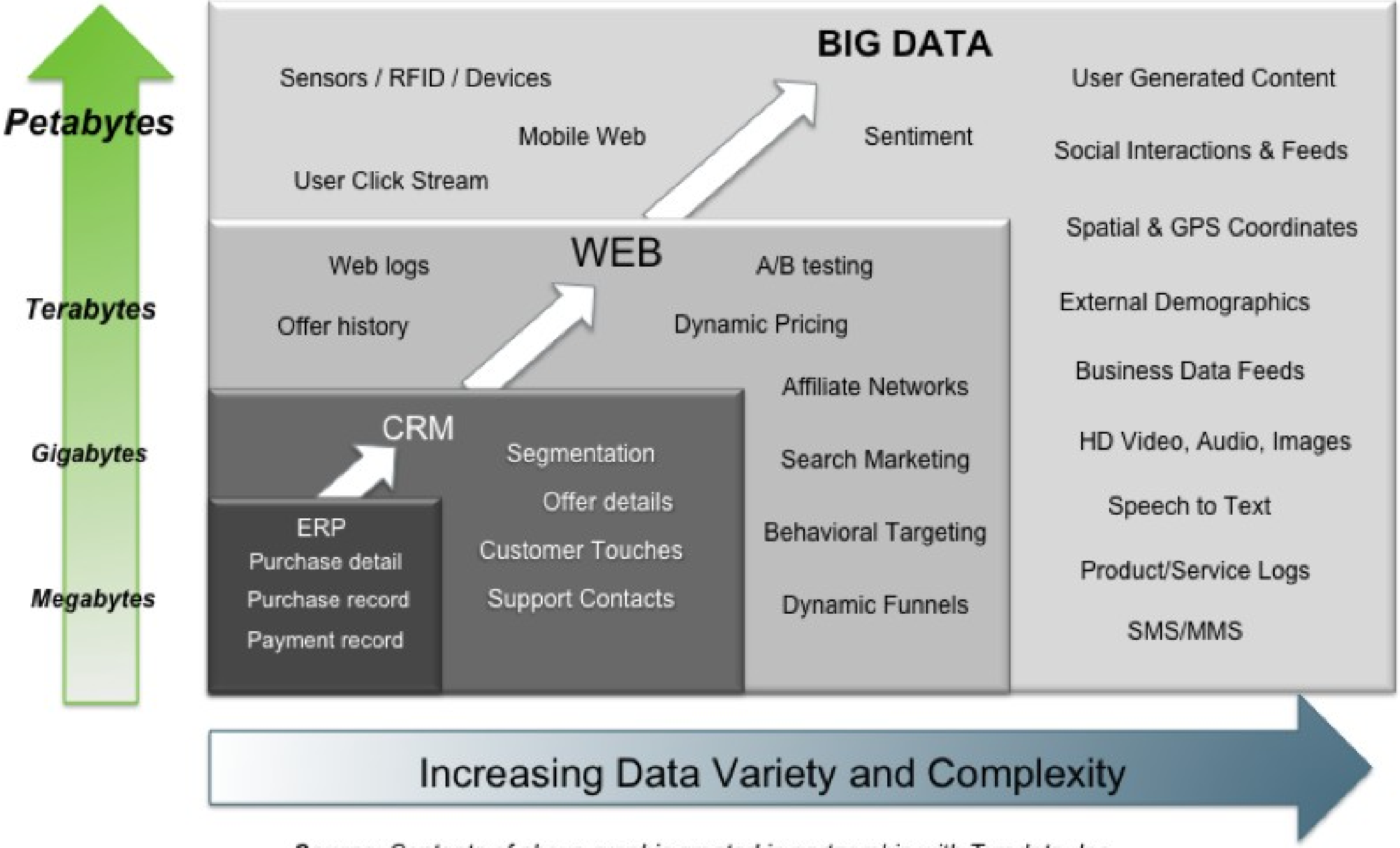
- Understand the organizational business (domain knowledge)
- Identify available data in an organization, and identify what potential value can be extracted out of the data and its correlation with other datasets (create hypothesis)
- Produce the methodologies and processes for analyzing the data to extract the value. (test the hypothesis).
  - In a Big Data environment, this have to be done through programming analysis code to instruct computers to analyze raw data for you.
- Create visualization and dashboards to present the analyzed data (provide a method for decision makers to digest the processed data and derive their inference from it)
- Continuously improve the data analysis code to improve its analysis quality and accuracy

Checkpoint

- What are the activities in data analysis process?
- Name 3 disciplines of Data Science

Evolution towards Big Data

Big Data = Transactions + Interactions + Observations



Petabytes

Terabytes

Gigabytes

Megabytes

BIG DATA

Sensors / RFID / Devices

Mobile Web

User Click Stream

Sentiment

User Generated Content

Social Interactions & Feeds

Spatial & GPS Coordinates

External Demographics

Business Data Feeds

HD Video, Audio, Images

Speech to Text

Product/Service Logs

SMS/MMS

WEB

Web logs

Offer history

A/B testing

Dynamic Pricing

Affiliate Networks

Search Marketing

Behavioral Targeting

Dynamic Funnels

CRM

Segmentation

Offer details

Customer Touches

Support Contacts

ERP

Purchase detail

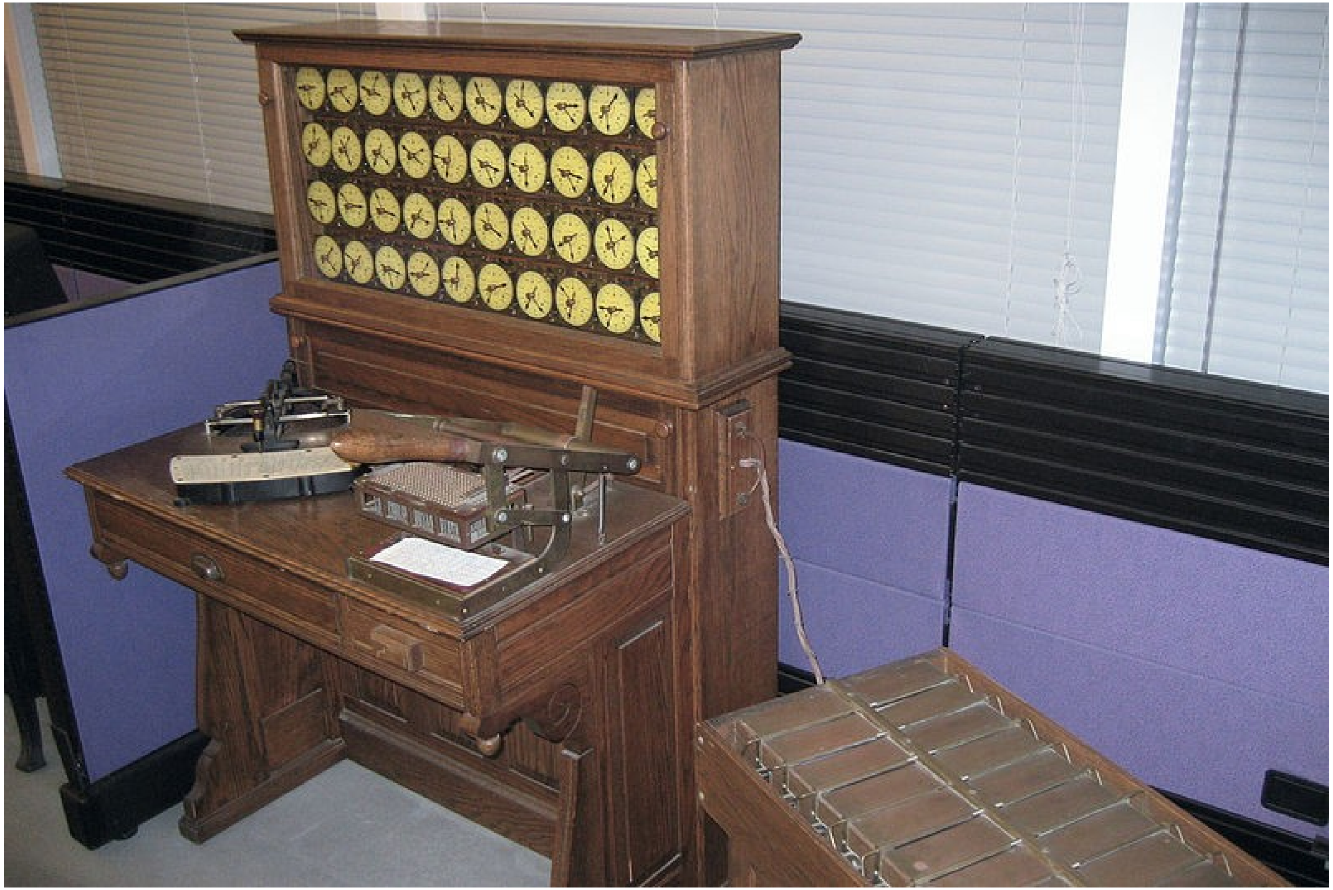
Purchase record

Payment record

Increasing Data Variety and Complexity

***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Pre Computer Era



Tabulating Machines

- 1890s – Herman Hollerith invented a mechanical device which records data onto punch cards, and calculate statistics of US Census
  - ".. finished months ahead of schedule and far under budget.. "
- 1900s – Tabulating machines adapted to aid in accounting and inventory
- Hollerith then founded a company to develop tabulating machines, which later become part of a merger that forms International Business Machines (IBM)

Early Data Analysis

- Enterprise Resource Planning (ERP) and accounting softwares gather business data and analytics are applied on the data to help improve businesses
  - Single database
  - Small datasets
  - Structured data only
  - Easy to analyze

Early Data Analysis

- Customer Relationship Management (CRM) softwares introduces more datasets and data types into the ecosystem
  - Unstructured data
    - Customer comments
    - Customer information
    - Contacts
  - More structured data sources

Dawn of the World Wide Web

- More generated data
  - Logs
  - System activity data
  - Network activity data
  - Marketing data
  - Advertising data
  - Etc, etc, etc

- Data still generated by organizational activities
- Data format not limited to databases
- Analysis complexity increases

Web2.0, Cloud and Software-as-a-Service

- User generated contents
  - Facebook posts
  - Tweets
  - Blogs
  - Articles
  - Instagram posts
  - Youtube Videos
  - Personal websites
  - User activities
  - Etc, etc, etc

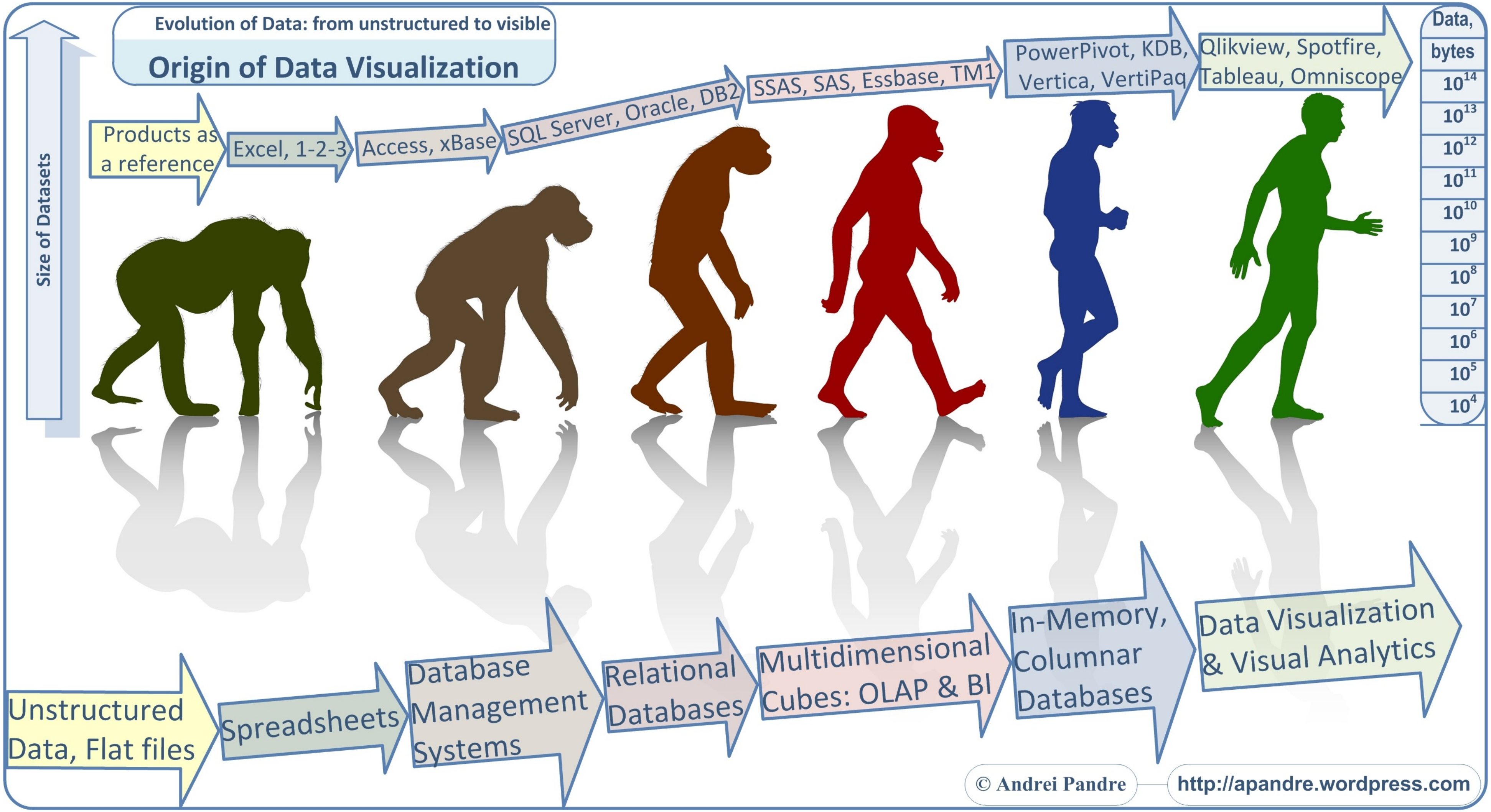
- User generated contents create an explosion in data growth
- The need for a new technologies to handle this data grow
- Hadoop, NoSQL and idea of Big Data was born

Internet Of Things

- The next step in data collecting
  - Sensors, sensors everywhere
  - High volume, high velocity data from various types of sensors
- Data from these sensors can be analyzed to gain insights and improve quality of life
  - Traffic information
  - Health information
  - Etc

Evolution of Data: from unstructured to visible

Origin of Data Visualization



Size of Datasets

Products as a reference → Excel, 1-2-3 → Access, xBase → SQL Server, Oracle, DB2 → SSAS, SAS, Essbase, TM1 → PowerPivot, KDB, Vertica, VertiPaq → Qlikview, Spotfire, Tableau, Omniscope

Unstructured Data, Flat files → Spreadsheets → Database Management Systems → Relational Databases → Multidimensional Cubes: OLAP & BI → In-Memory, Columnar Databases → Data Visualization & Visual Analytics

Data, bytes
10^{14}
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}

© Andrei Pandre — http://apandre.wordpress.com

Checkpoint

- What was the earliest example of machine aided data analysis?
- What was the name/type of the machine?
- Why does the move towards Web 2.0, Cloud and SaaS cause an explosion of raw data?

Business Intelligence
Data Warehouse

Definition

Business intelligence (BI) is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities.

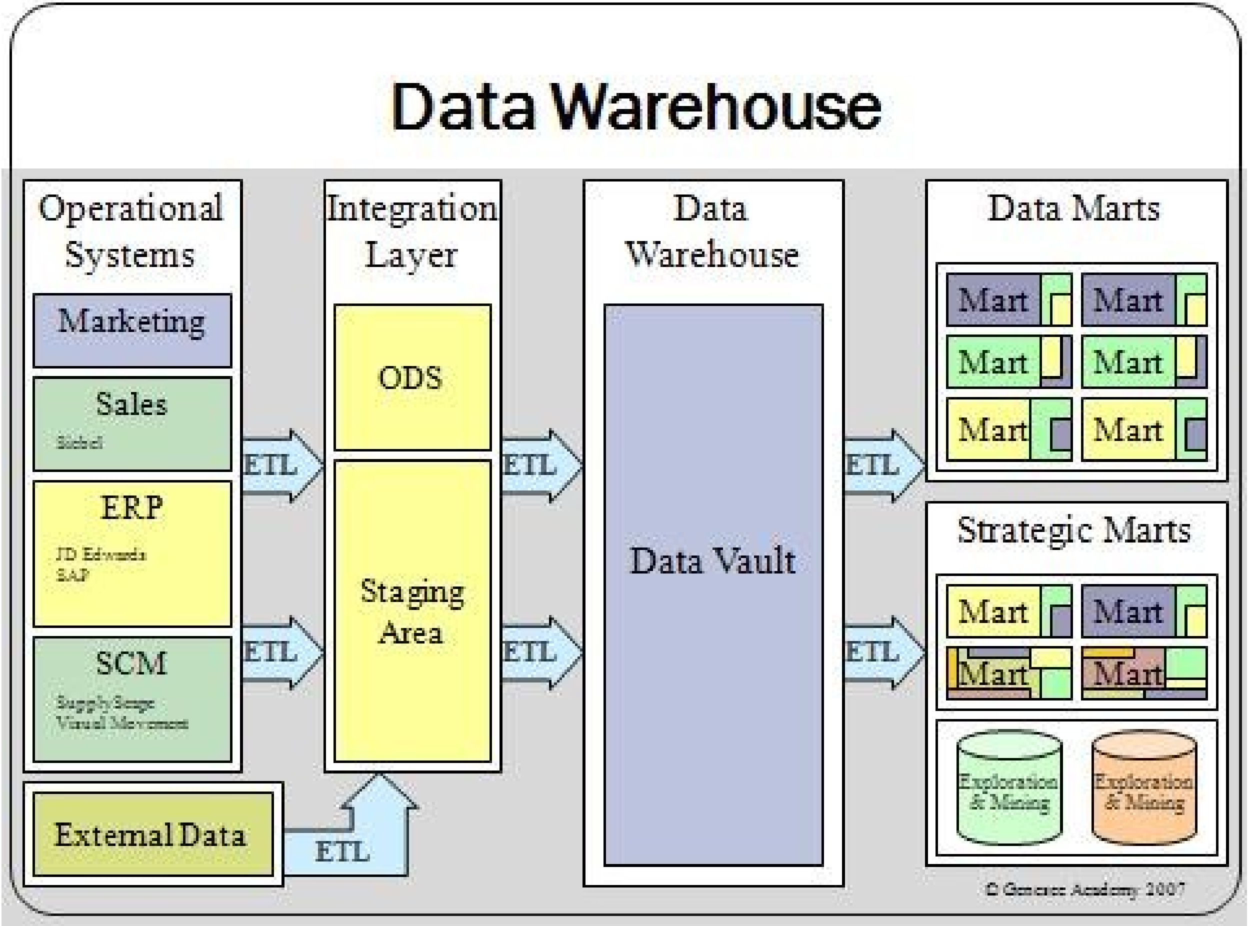
Components of a BI tool

- Reporting
- Online analytical processing
- Analytics
- Data mining
- Process mining
- Complex event processing
- Business performance management
- Benchmarking
- Text mining
- Predictive analytics
- Prescriptive analytics

Data Warehouse

- Business Intelligence tools analyze data
- Data warehouse role is to integrate, store and chunk the data to make them more accessible for BI tools to analyze
- ETL – Extract Transform Load handles the process of analysis

Data Warehouse Architecture



Components of a Data Warehouse

- Data Sources
  - Existing data from operational databases, flat files, logs, etc
  - External data from external sources
- ETL – Extract Transform Load (Layer 1)
  - Massage and clean the data to prepare for storage inside data warehouse
  - Integrate common datasets into a more digestible formats
- Data Warehouse
  - Centralized repository for storing all available datasets for it to be easily accessible for modeling and analysis

Components of a Data Warehouse

- ETL – Extract Transform Load (Layer 2)
  - Model data into datamarts to make them easily analyzed by Business Intelligence tools
  - Chunk data into smaller datasets for sharing with other organizations
- Data Marts
  - A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse that is usually oriented to a specific business line or team. Data marts are small slices of the data warehouse.
  - Data in data marts are provided in format that can be easily consumed by other softwares and users

Data Analysis on BIDW

- ETL process handles data cleansing and transformation of data
- Business Intelligence software analyzes statistics of data from data marts
- Business intelligence software then is used to build reports and dashboards to present the results

Checkpoint

- Name 3 components of a Business Intelligence software
- Name 3 components of a Data Warehouse
- What is the importance of a Data Mart?

Big Data Analytics

VVVV of Big Data

Volume
Velocity
Variety
Variability

Challenge of Big Data

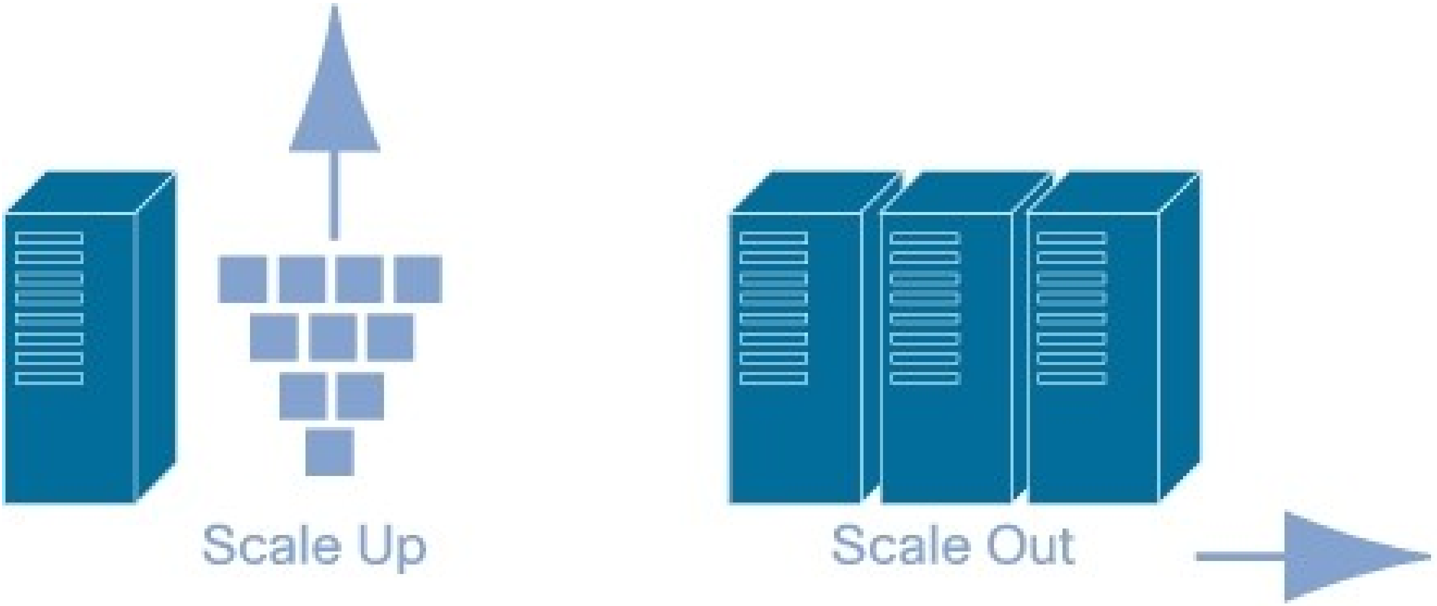
- Volume – size of data
  - Data size too large for BIDW too analyze
  - Petabytes of data can't be stored in a single data warehouse
  - Traditional BIDW architecture too slow to query and analyze

Challenge of Big Data

- Velocity – speed of data coming in
  - Link back to volume – data grow at a very fast rate – storage run out fast
  - Need to add capacity in a rapid, horizontal manner
  - Traditional data warehouse scales vertically instead of horizontally – database sharding and partitioning is complex and could not scale rapid enough

Pit Stop

- What is Vertical vs Horizontal Scaling?
  - **To scale vertically (or scale up)** means to add resources to a single node in a system, typically involving the addition of CPUs, memory or storage to a single computer.
  - **To scale horizontally (or scale out)** means to add more nodes to a system, such as adding a new computer to a distributed software application


Scale Up
Scale Out

Challenge of Big Data

- Variety – various data types and formats
  - Semi structured data
    - Logs, sensors
  - Unstructured data
    - Opinion, words and sentence, documents, videos, audio, images
  - Many traditional BIDW focus on structured data, with limited capability to analyze unstructured data

Challenge of Big Data

- Variability – data that keep changing or of complex types
  - Inconsistent and erratic data sources
  - Require neural logic to construct and model the data sources into meaningful format
  - Data from diverse security models, application interfaces, metadata schemas, 3rd party data, etc

Hadoop and Big Data Challenges

Origins of Hadoop

- 2002 - Mike Cafarella and Doug Cutting started a project to create an open source web crawler and search engine - Nutch
- 2003 – Google publishes a paper about their distributed file system (GFS)
  - Describes how they solve their storage needs in a massive search engine environment
  - Mike and Doug started the Nutch Distributed File System project on 2004
- 2004 – Google publishes a paper that introduces MapReduce to the world
  - Describes how Google utilizes the method to scale out their datacenter processing
  - By 2005, Nutch gets its own MapReduce implementation
- February 2006 – NDFS and MapReduce implementation in Nutch were applicable for things beyond search engine – a new project was born – Hadoop

Hadoop as a Solution to Big Data Challenges

- Volume
  - Hadoop Distributed File System solve the problem of volume through horizontal scalability
  - Capacity can be increased in a rapid manner, and built-in redundancy provides resilient storage of data even on hardware failure
- Velocity
  - Hadoop data loading tools are designed to be executed in a distributed manner, allowing data to be fed and stored into a Dadoop cluster in high velocity

Hadoop as a Solution to Big Data Challenges

- Variety
  - MapReduce concept enables analysis of data beyond traditional structured datasets
  - Enable complex analysis processes to be executed against unstructured data in highly scalable and distributed manner
- Variability
  - MapReduce functions can be developed with sophisticated logic to handle variability of data
  - Besides MapReduce, Hadoop distributions also includes Machine Learning and Data Mining tools, which enables a method to teach the system to handle variable datasets

Do I Need Hadoop?

- From the VVVV of Big Data, any of them applies to you in your data analytics needs?
  - If yes, you probably need it
  - If no, you probably don't,
    - But you might in the future
    - Or might not
  - Subjective, and case by case basis
- If what you are looking for is a way for organizations/agencies to share data with each other for their own analytics needs – you are looking for Open Data, not necessarily Big Data.

Checkpoint

- What are the challenges introduced by the VVVV of Big Data?
- How Hadoop aid in solving the challenges?

Activity

- Split into groups of 8 people
- Identify what are the data you may have in your respective organizations
  - Structured? Unstructured? Sensors? Logs?
- Identify what type of analysis can be done on the datasets, identify what are the value that can be extracted out from the datasets
- Identify whether the datasets can be analyzed with just traditional Business Intelligence, or do you need a Hadoop / Big Data infrastructure in place to process the analysis in less than 12 hours.
  - If you don't need Big Data yet, do you foresee the need to come soon in the future? (eg: introduction of more data gathering activities such as sensors in the organization)
- Discuss with your team for 30 minutes, and present your findings

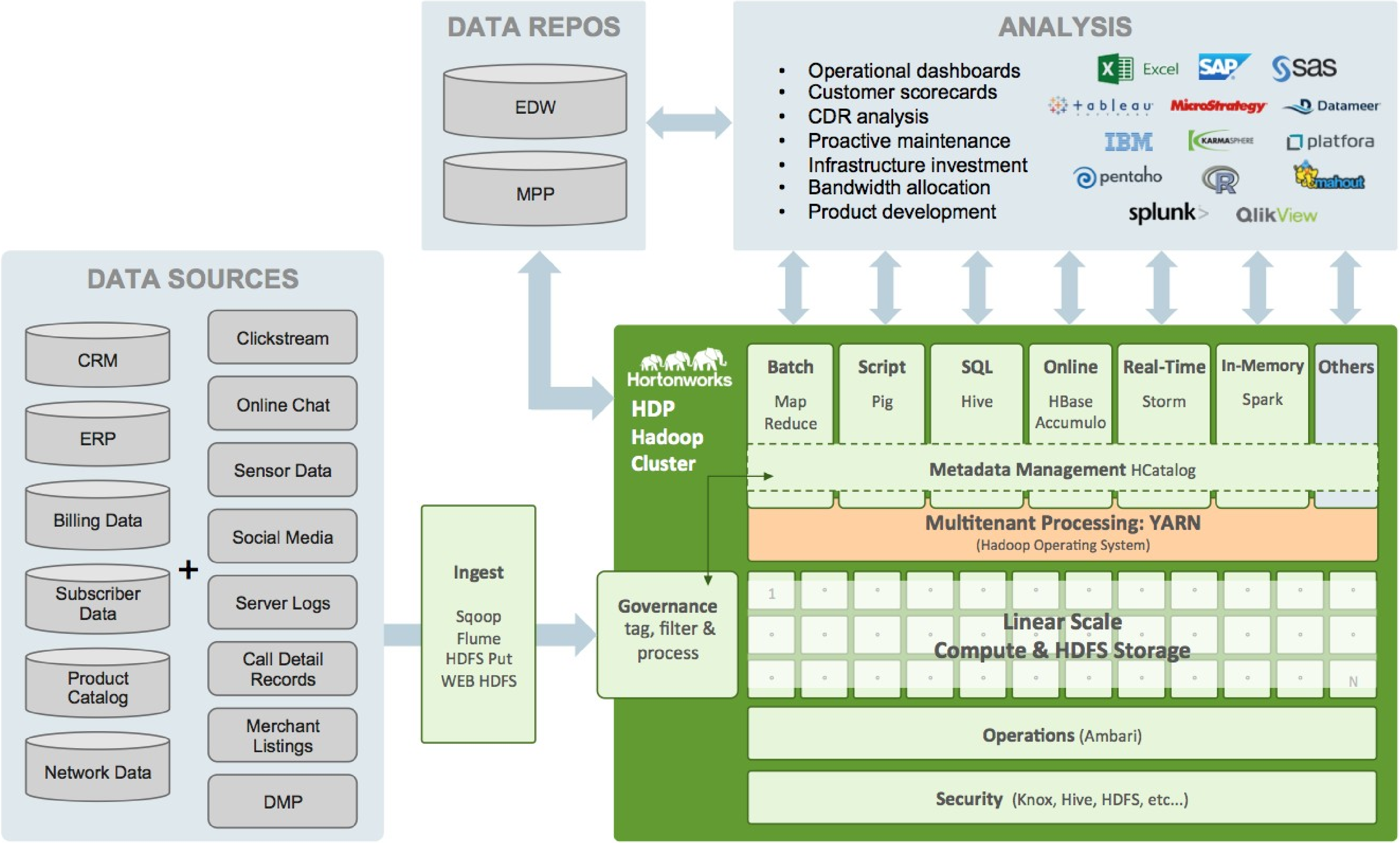
Building a Modern Data Architecture

Big Data Architecture

- Data Lake Architecture
  - A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data.
  - HDFS in Hadoop allows creation of Data Lake

Big Data Architecture

- Data Refinery Architecture
  - A data refinery is a facility for transforming raw data into relevant and actionable information. Data refinement services take the uncertainty out of the data foundation for analysis and operations. Refined data is timely, clean and well understood.
  - HDFS + Hadoop analysis tools when used together creates a Data Refinery



DATA REPOS

EDW

MPP

ANALYSIS

- Operational dashboards
- Customer scorecards
- CDR analysis
- Proactive maintenance
- Infrastructure investment
- Bandwidth allocation
- Product development

Excel SAP sas tableau MicroStrategy Datameer IBM KARMASPHERE platfora pentaho R mahout splunk> QlikView

DATA SOURCES

CRM

ERP

Billing Data

Subscriber Data

Product Catalog

Network Data

+

Clickstream

Online Chat

Sensor Data

Social Media

Server Logs

Call Detail Records

Merchant Listings

DMP

Ingest

Sqoop
Flume
HDFS Put
WEB HDFS

Hortonworks

HDP Hadoop Cluster

Batch Map Reduce

Script Pig

SQL Hive

Online HBase Accumulo

Real-Time Storm

In-Memory Spark

Others

Metadata Management HCatalog

Multitenant Processing: YARN
(Hadoop Operating System)

Governance tag, filter & process

1

N

Linear Scale Compute & HDFS Storage

Operations (Ambari)

Security (Knox, Hive, HDFS, etc...)

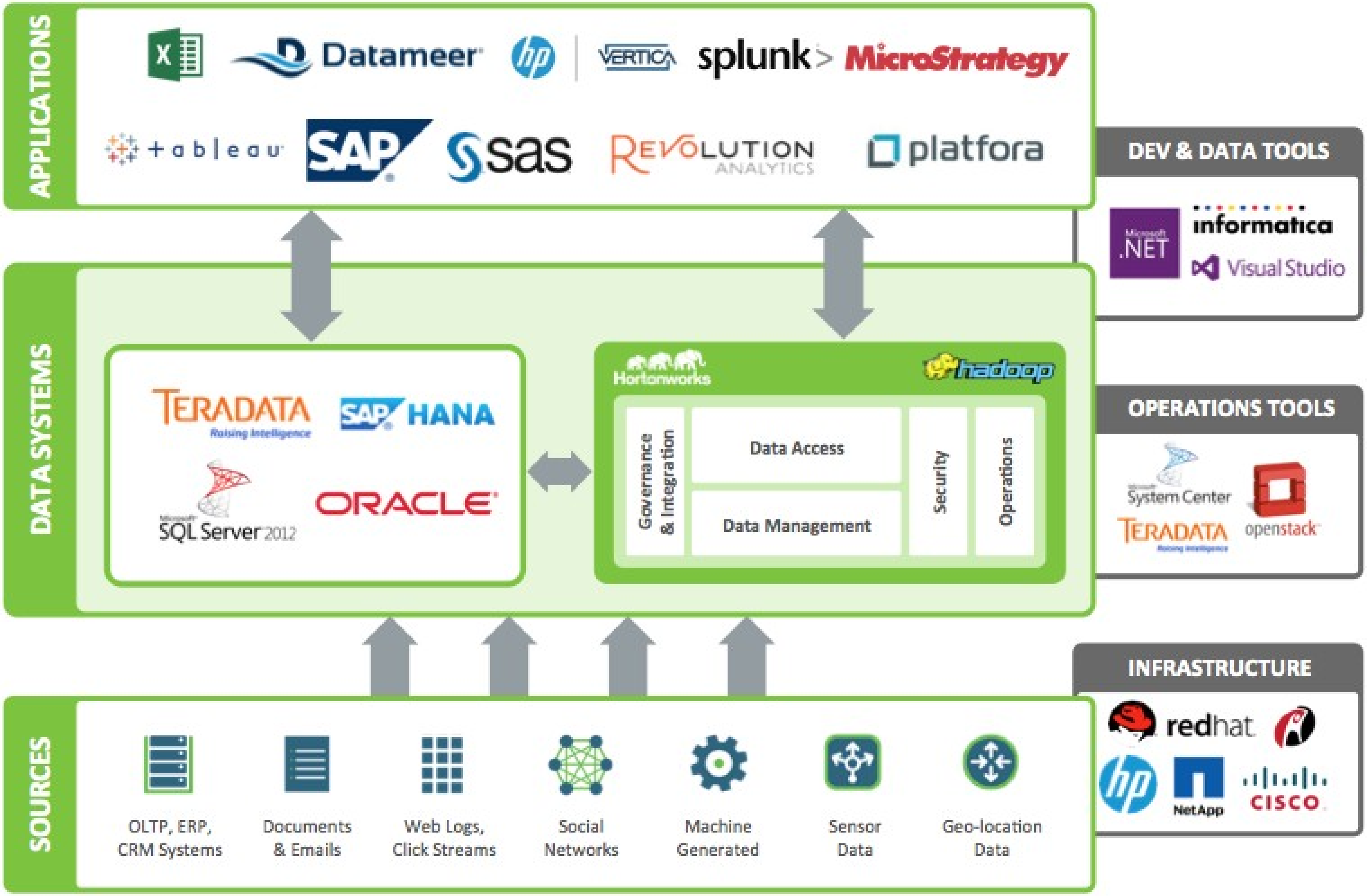
Hadoop alongside Traditional Data Architecture

Hybrid Data Warehouse and Data Lake/Refinery Architecture

- Traditional BIDW are excellent for the type of workloads it does best
- Current Hadoop technology is not quite there yet to replace traditional Data Warehouse – Relative to DW, MapReduce is slow when handling small datasets
  - HiveQL performance improvements are continuously being developed – Recent switch from MapReduce to Tez improve realtime performance
  - Apache Spark (built on top of Hadoop) aims to bring 100x faster performance on certain applications
- Organizations that have invested on data warehousing can continue utilize DW to handle traditional workload, and introduce Hadoop as a plug-in into the architecture to handle Big Data workload

Hybrid Data Warehouse and Data Lake/Refinery Architecture

- Traditional BIDW are excellent for the type of workloads it does best
- Current Hadoop technology is not quite there yet to replace traditional Data Warehouse – Relative to DW, MapReduce is slow when handling small datasets
  - HiveQL performance improvements are continuously being developed – Recent switch from MapReduce to Tez improve realtime performance
  - Apache Spark (built on top of Hadoop) aims to bring 100x faster performance on certain applications
- Organizations that have invested on data warehousing can continue utilize DW to handle traditional workload, and introduce Hadoop as a plug-in into the architecture to handle Big Data workload



APPLICATIONS

DATA SYSTEMS

Governance & Integration

Data Access

Data Management

Security

Operations

SOURCES

OLTP, ERP, CRM Systems

Documents & Emails

Web Logs, Click Streams

Social Networks

Machine Generated

Sensor Data

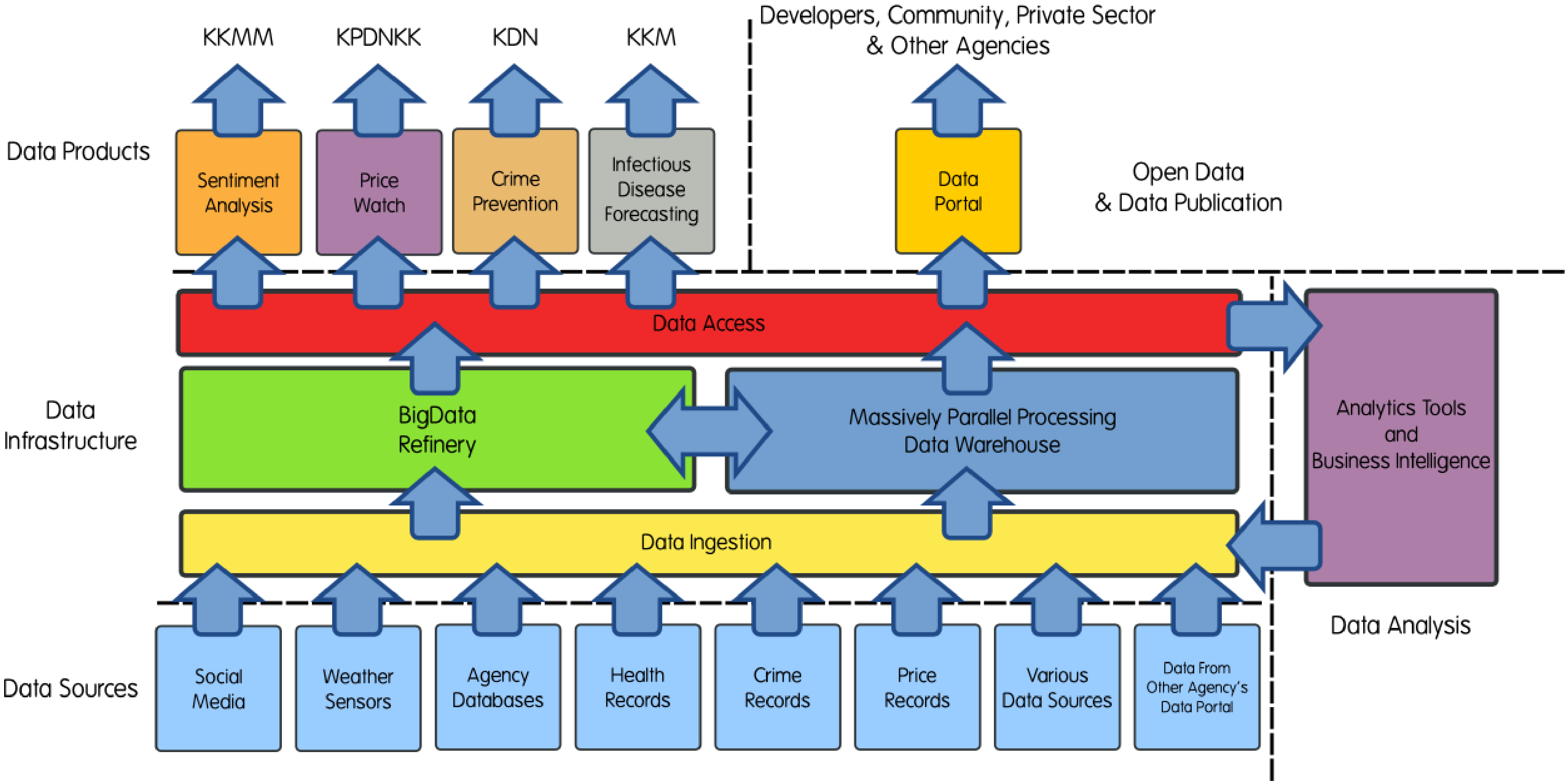
Geo-location Data

DEV & DATA TOOLS

OPERATIONS TOOLS

INFRASTRUCTURE

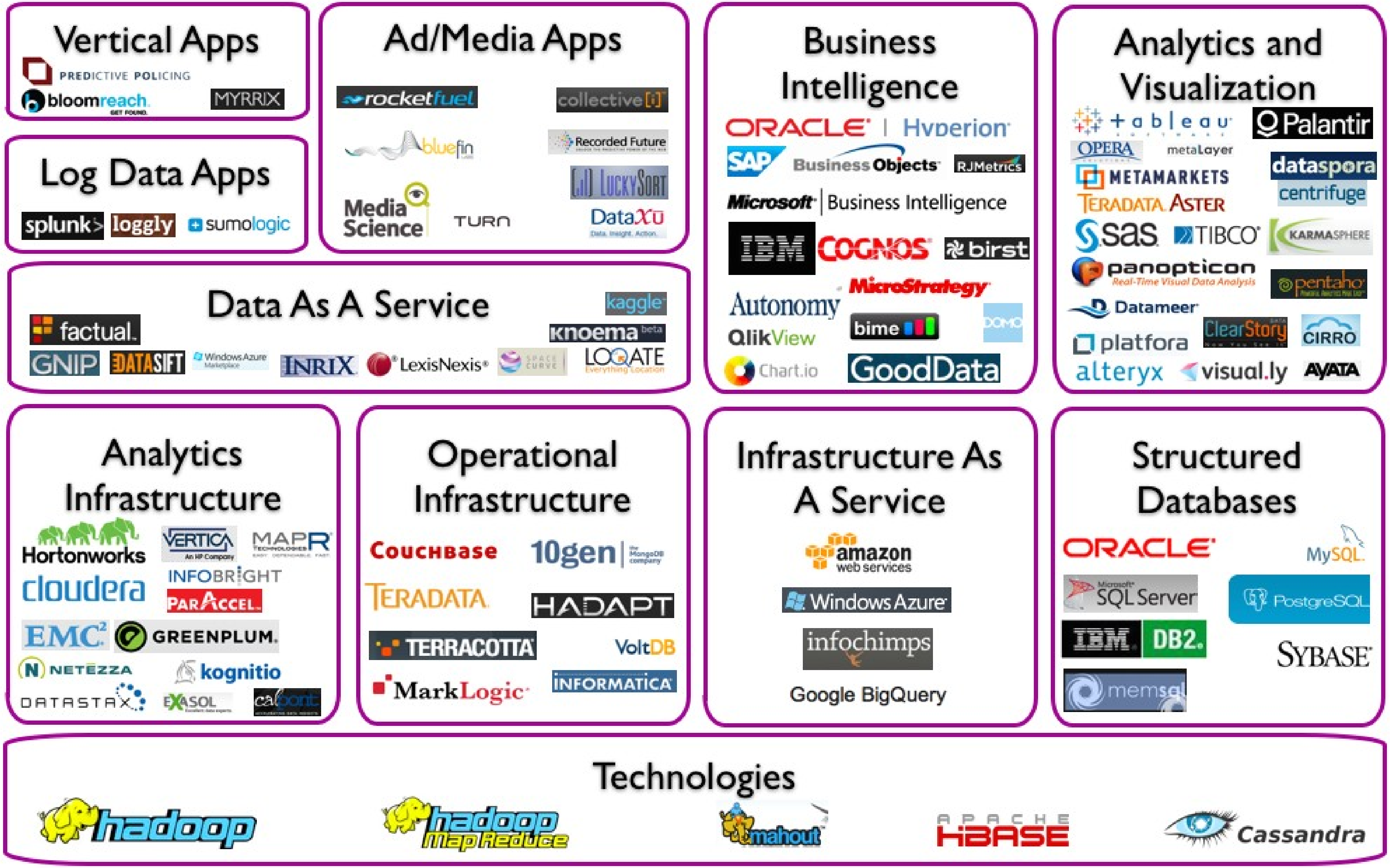
Data Architecture In Context of Malaysian Government Initiatives: An example of possible implementation


KKMM
KPDNKK
KDN
KKM
Developers, Community, Private Sector & Other Agencies
Data Products
Sentiment Analysis
Price Watch
Crime Prevention
Infectious Disease Forecasting
Data Portal
Open Data & Data Publication
Data Access
Data Infrastructure
BigData Refinery
Massively Parallel Processing Data Warehouse
Analytics Tools and Business Intelligence
Data Ingestion
Data Analysis
Data Sources
Social Media
Weather Sensors
Agency Databases
Health Records
Crime Records
Price Records
Various Data Sources
Data From Other Agency's Data Portal

Architecture Overview

- Data Ingestion
  - Loads and integrate datasets from various sources and types
- Data Warehouse
  - Handles traditional BIDW workload
- Data Refinery
  - Big Data analytics runs here
- Data Products
  - The 4 government pioneer projects are technically data products
- Data Sharing / Open Data
  - Publication of data for 3rd party organizations, businesses, and developers to utilize the data in their own data analysis purposes
- Data Analysis and Business Intelligence
  - Provides the tools, technologies and expertise (human resource) in extracting value and analytics out of stored data

Software Products in Big Data Ecosystem



Vertical Apps

PREDICTIVE POLICING
bloomreach GET FOUND.
MYRRIX

Log Data Apps

splunk>
loggly
sumologic

Ad/Media Apps

rocketfuel
collective[i]
bluefin LABS
Recorded Future
LuckySort
Media Science
TURN
DataXu Data. Insight. Action.

Data As A Service

factual.
kaggle
knoema beta
GNIP
DATASIFT
Windows Azure Marketplace
INRIX
LexisNexis
SPACE CURVE
LOQATE Everything Location

Business Intelligence

ORACLE | Hyperion
SAP
Business Objects
RJMetrics
Microsoft | Business Intelligence
IBM
COGNOS
birst
MicroStrategy
Autonomy
bime
DOMO
QlikView
Chart.io
GoodData

Analytics and Visualization

tableau software
Palantir
OPERA SOLUTIONS
metaLayer
dataspora
METAMARKETS
centrifuge
TERADATA ASTER
sas
TIBCO
KARMASPHERE
panopticon Real-Time Visual Data Analysis
pentaho
Datameer
ClearStory
CIRRO
platfora
alteryx
visual.ly
AYATA

Analytics Infrastructure

Hortonworks
VERTICA An HP Company
MAPR TECHNOLOGIES
cloudera
INFOBRIGHT
PARACCEL
EMC²
GREENPLUM.
NETEZZA
kognitio
DATASTAX
EXASOL
calpont

Operational Infrastructure

COUCHBASE
10gen | The MongoDB company
TERADATA
HADAPT
TERRACOTTA
VoltDB
MarkLogic
INFORMATICA

Infrastructure As A Service

amazon web services
Windows Azure
infochimps
Google BigQuery

Structured Databases

ORACLE
MySQL
Microsoft SQL Server
PostgreSQL
IBM DB2.
SYBASE
memsql

Technologies

hadoop
hadoop Map Reduce
mahout
APACHE HBASE
Cassandra

End of Day 1