



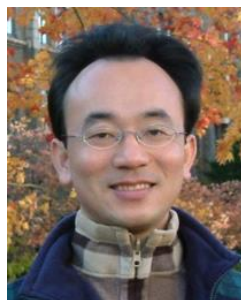
Data Mining 資料探勘

海量資料分析 (Big Data Analytics)

1032DM09

MI4

Wed, 7,8 (14:10-16:00) (B130)



Min-Yuh Day

戴敏育

Assistant Professor

專任助理教授

Dept. of Information Management, Tamkang University

淡江大學 資訊管理學系

<http://mail.tku.edu.tw/myday/>

2015-05-13



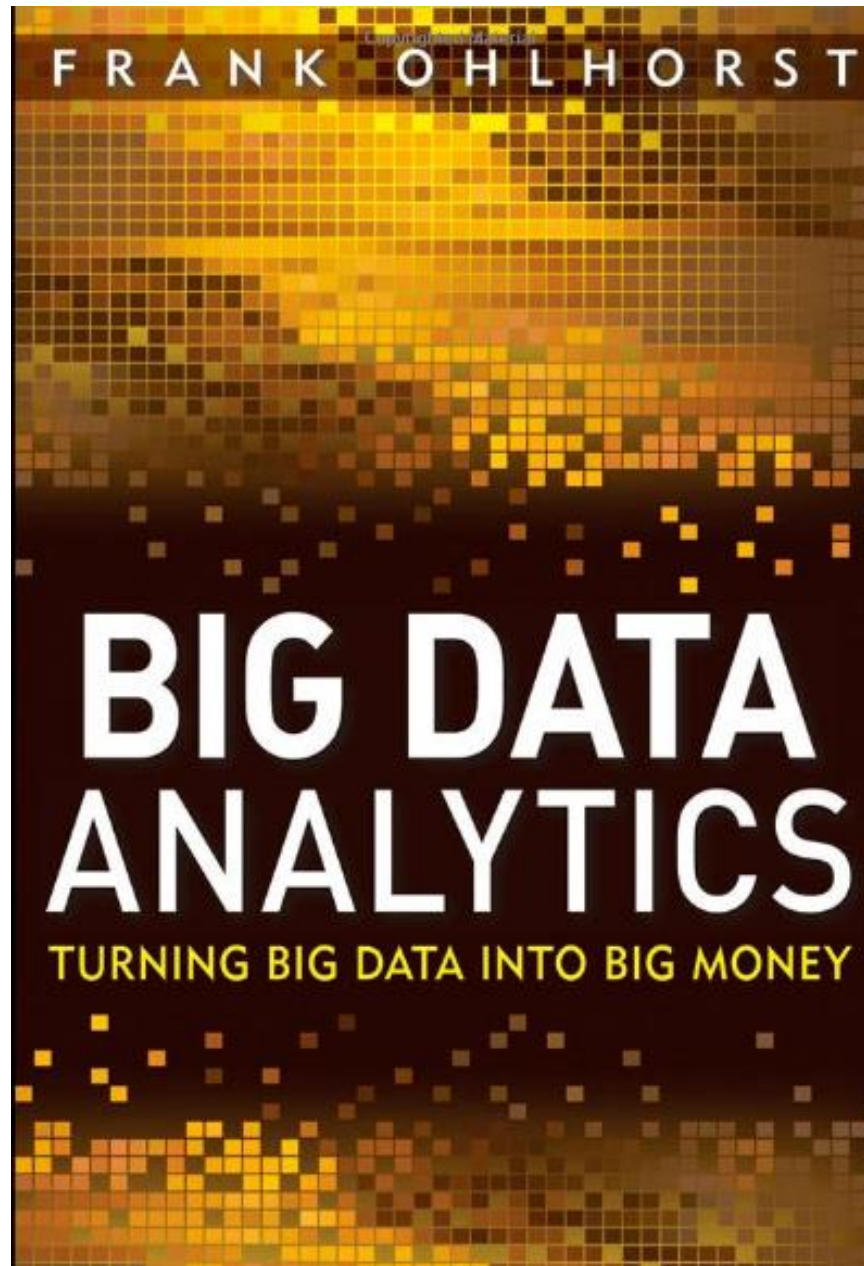
課程大綱 (Syllabus)

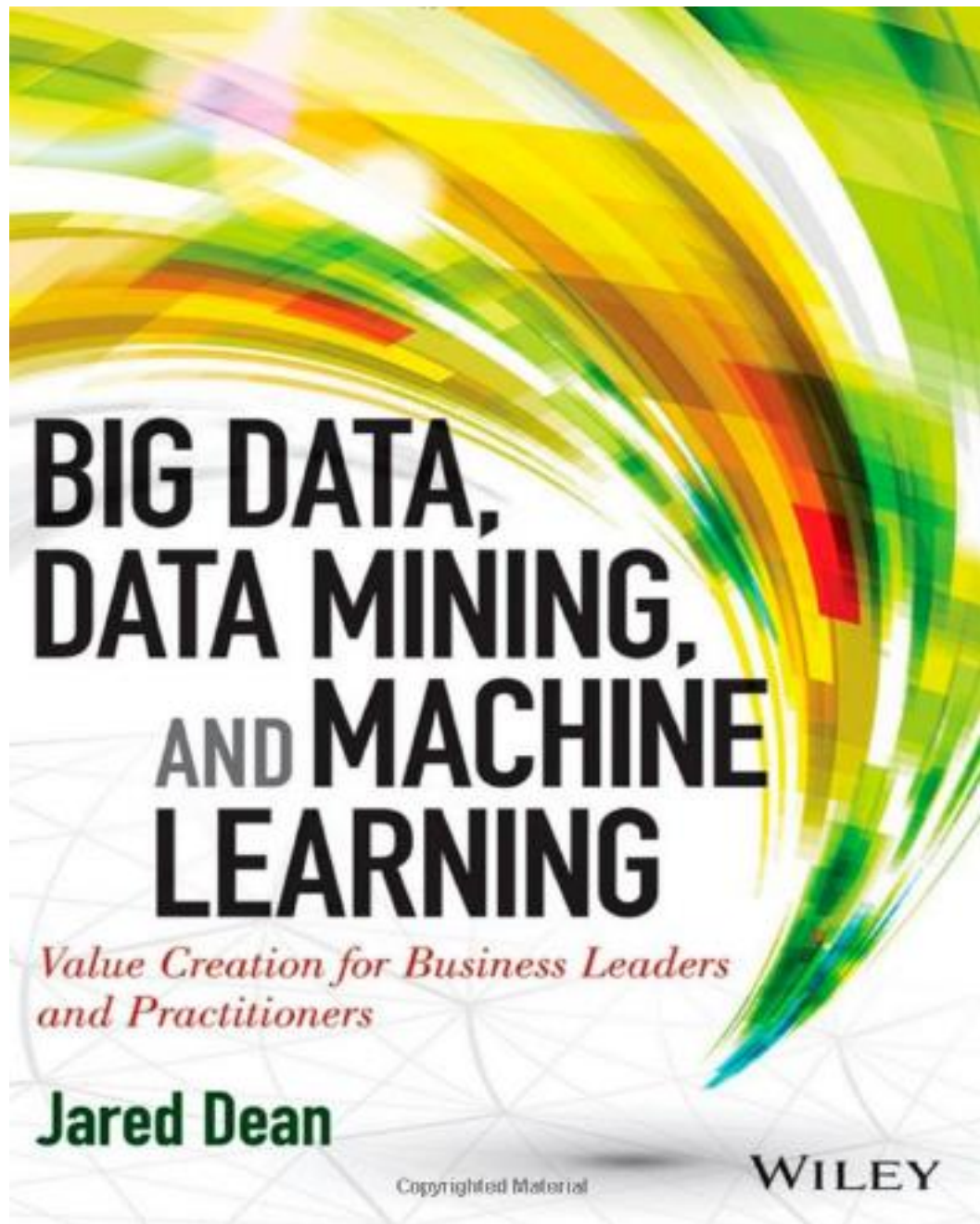
週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2015/02/25	資料探勘導論 (Introduction to Data Mining)
2	2015/03/04	關連分析 (Association Analysis)
3	2015/03/11	分類與預測 (Classification and Prediction)
4	2015/03/18	分群分析 (Cluster Analysis)
5	2015/03/25	個案分析與實作一 (SAS EM 分群分析) : Case Study 1 (Cluster Analysis – K-Means using SAS EM)
6	2015/04/01	教學行政觀摩日 (Off-campus study)
7	2015/04/08	個案分析與實作二 (SAS EM 關連分析) : Case Study 2 (Association Analysis using SAS EM)
8	2015/04/15	個案分析與實作三 (SAS EM 決策樹、模型評估) : Case Study 3 (Decision Tree, Model Evaluation using SAS EM)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
9	2015/04/22	期中報告 (Midterm Project Presentation)
10	2015/04/29	期中考試週 (Midterm Exam)
11	2015/05/06	個案分析與實作四 (SAS EM 迴歸分析、類神經網路) : Case Study 4 (Regression Analysis, Artificial Neural Network using SAS EM)
12	2015/05/13	海量資料分析 (Big Data Analytics)
13	2015/05/20	文字探勘與網頁探勘 (Text and Web Mining)
14	2015/05/27	期末報告 (Final Project Presentation)
15	2015/06/03	畢業考試週 (Final Exam)

Big Data Analytics





BIG DATA, DATA MINING, AND MACHINE LEARNING

*Value Creation for Business Leaders
and Practitioners*

Jared Dean

WILEY

Copyrighted Material

PREDICTIVE ANALYTICS

AN INTRODUCTION
FOR EVERYONE



THE POWER TO PREDICT WHO WILL
CLICK, BUY, LIE, OR DIE

ERIC SIEGEL



Source: <http://www.amazon.com/Big-Data-Revolution-Transform-Mayer-Schonberger/dp/B00D81X2YE>

Wiley CIO Series

Copyrighted Material

Foreword by
JIM STOGDILL
General Manager
Radar,
O'Reilly Media

BIG DATA BIG ANALYTICS

EMERGING BUSINESS INTELLIGENCE AND
ANALYTIC TRENDS FOR TODAY'S
BUSINESSES

Michael Minelli • Michele Chambers • Ambiga Dhiraj

Copyrighted Material

Big Data, Big Analytics:

**Emerging Business Intelligence
and Analytic Trends
for Today's Businesses**

Big Data, Prediction vs. Explanation

Big Data:

The Management Revolution

HBR.ORG

Harvard Business Review



OCTOBER 2012
REPRINT R1210C

SPOTLIGHT ON BIG DATA

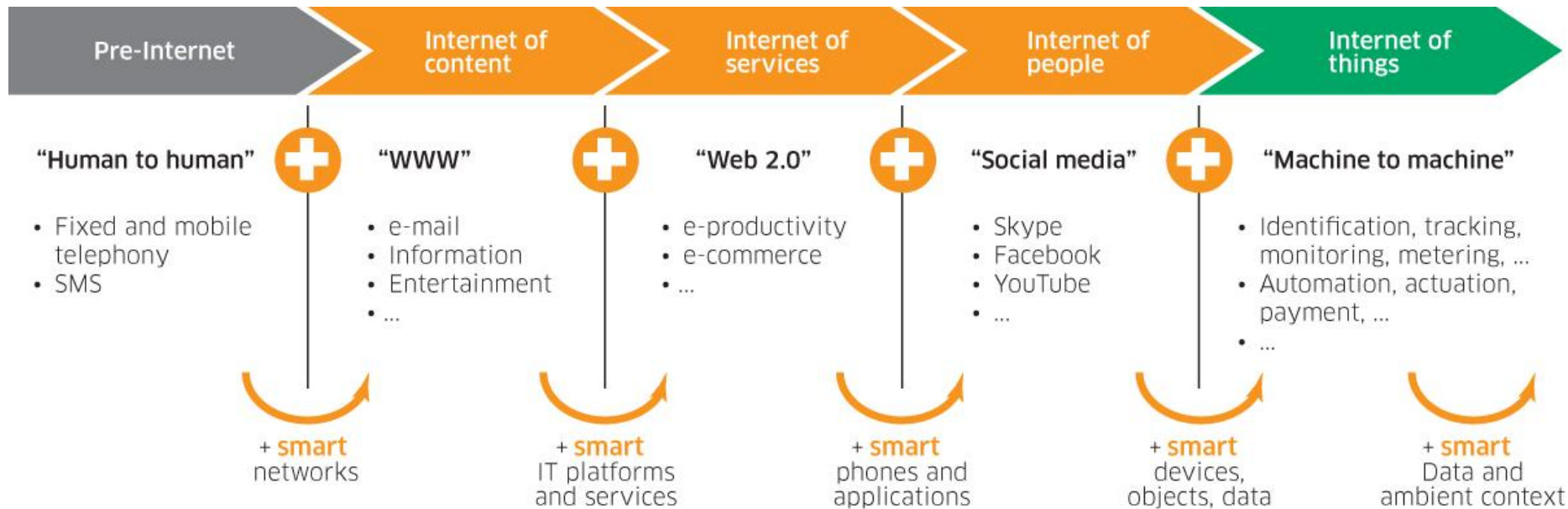
Big Data: The Management Revolution

Exploiting vast new flows of information can radically improve your company's performance. But first you'll have to change your decision-making culture.
by Andrew McAfee and Erik Brynjolfsson

Internet Evolution

Internet of People (IoP): Social Media

Internet of Things (IoT): Machine to Machine



Source: Marc Jadoul (2015), The IoT: The next step in internet evolution, March 11, 2015

<http://www2.alcatel-lucent.com/techzine/iot-internet-of-things-next-step-evolution/>

Copyrighted Material

ENTERPRISE ANALYTICS

Optimize Performance, Process, and
Decisions through Big Data



EDITED BY
THOMAS DAVENPORT

Copyrighted Material

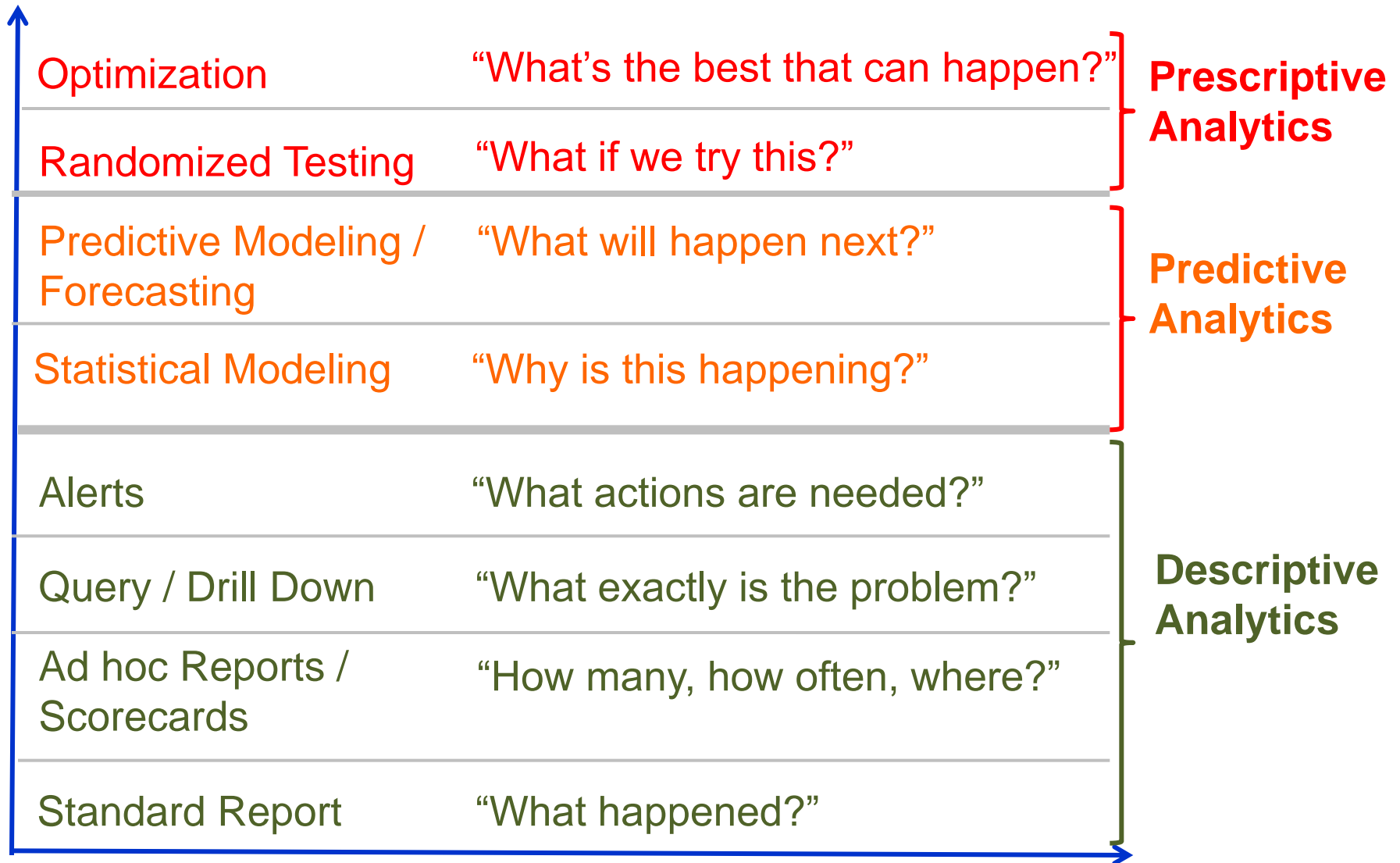
Business Intelligence and Enterprise Analytics

- Predictive analytics
- Data mining
- Business analytics
- Web analytics
- **Big-data** analytics

Three Types of Business Analytics

- Prescriptive Analytics
- Predictive Analytics
- Descriptive Analytics

Three Types of Business Analytics



Big-Data Analysis

- **Too Big,
too Unstructured,
too many different source
to be manageable through traditional
databases**

Business Intelligence and Analytics: Research Directions

1. Big Data Analytics

- Data analytics using Hadoop / MapReduce framework

2. Text Analytics

- From Information Extraction to Question Answering
- From Sentiment Analysis to Opinion Mining

3. Network Analysis

- Link mining
- Community Detection
- Social Recommendation

The Rise of “Big Data”

- “Too Big” means databases or data flows in **petabytes (1,000 terabytes)**
 - Google processes about 24 petabytes of data per day
- “Too unstructured” means that the data isn’t easily put into the traditional rows and columns of conventional databases

Examples of Big Data

- Online information
 - Clickstream data from Web and social media content
 - Tweets
 - Blogs
 - Wall postings
- Video data
 - Retail and crime/intelligence environments
 - Rendering of video entertainment
- Voice data
 - call centers and intelligence intervention
- Life sciences
 - Genomic and proteomic data from biological research and medicine

Big Data, Big Analytics:

Emerging Business Intelligence and Analytic Trends for Today's Businesses

- What Big Data is and why it's important
- Industry examples (Financial Services, Healthcare, etc.)
- Big Data and the New School of Marketing
- Fraud, risk, and Big Data
- Big Data technology
- Old versus new approaches
- Open source technology for Big Data analytics
- The Cloud and Big Data

Big Data, Big Analytics:

Emerging Business Intelligence and Analytic Trends for Today's Businesses

- Predictive analytics
- Crowdsourcing analytics
- Computing platforms, limitations, and emerging technologies
- Consumption of analytics
- Data visualization as a way to take immediate action
- Moving from beyond the tools to analytic applications
- Creating a culture that nurtures decision science talent
- A thorough summary of ethical and privacy issues

What is **BIG Data**?

Volume

Large amount of data

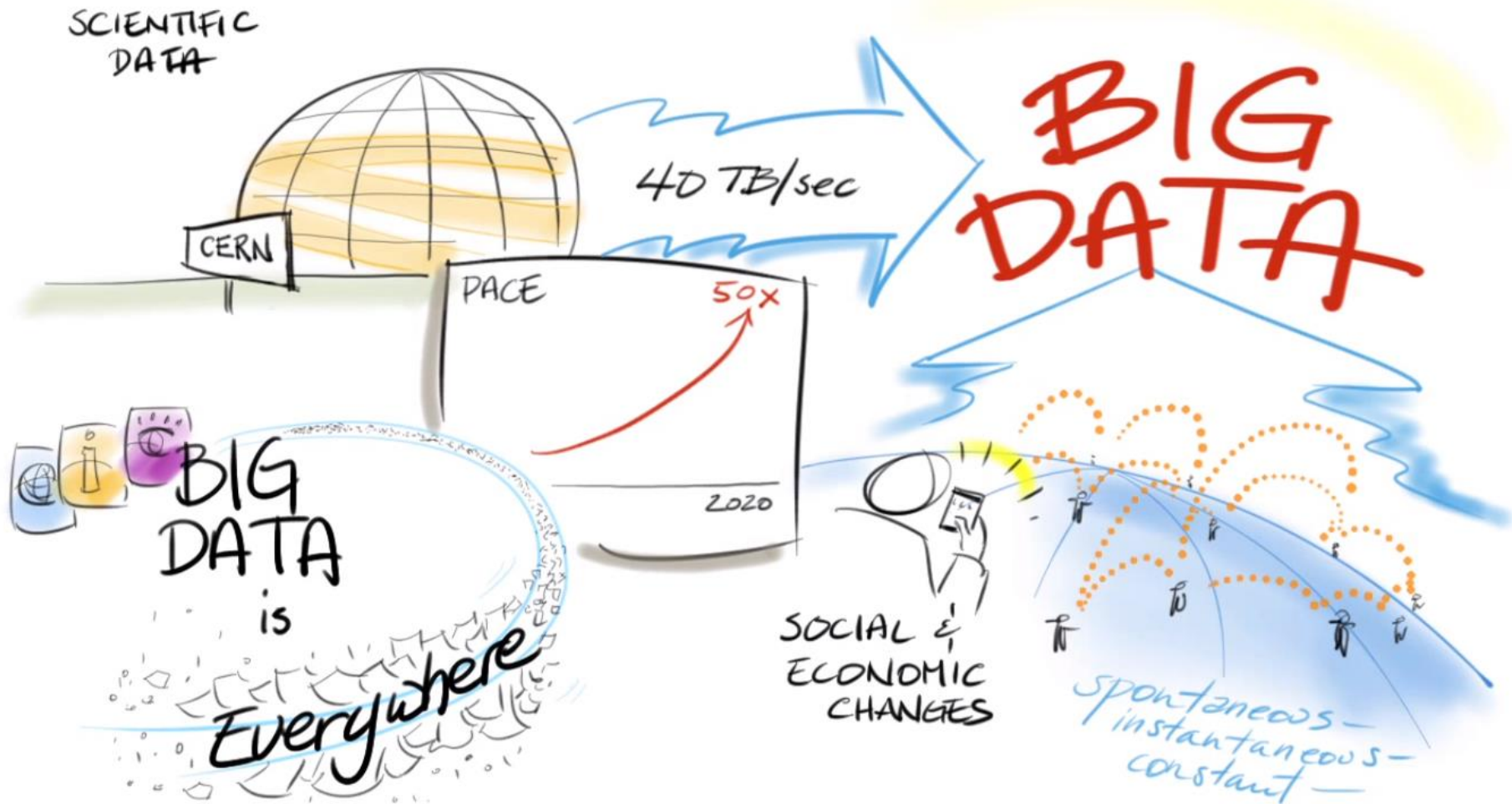
Velocity

Needs to be analyzed **quickly**

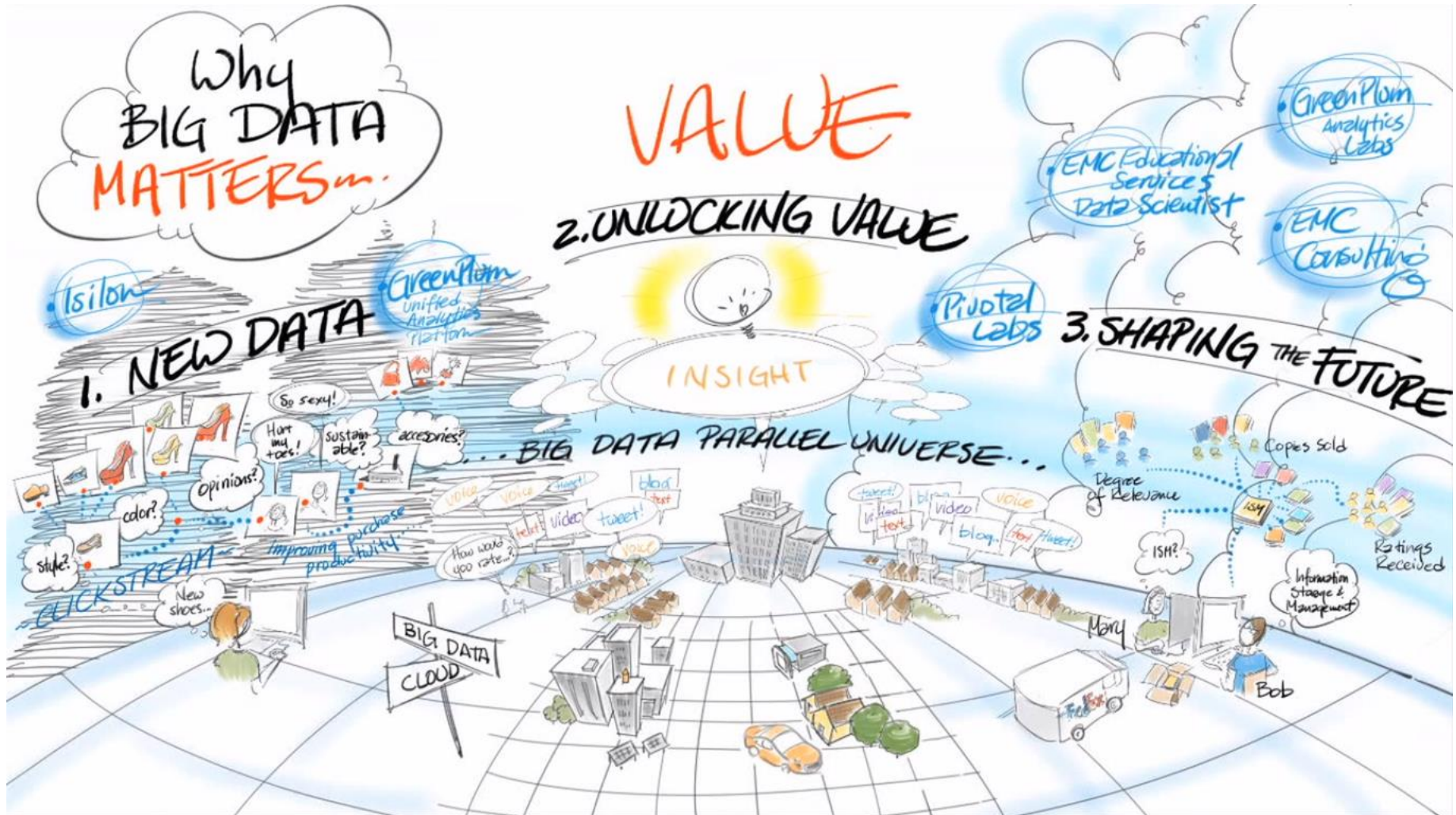
Variety

Different types of structured and unstructured data

Big Ideas: How Big is **Big Data**?



Big Ideas: Why **Big Data** Matters



Key questions enterprises are asking about **Big Data**

- How to store and protect big data?
- How to backup and restore big data?
- How to organize and catalog the data that you have backed up?
- How to keep costs low while ensuring that all the critical data is available when you need it?

Volumes of Data

- Facebook
 - **30 billion pieces of content** were added to Facebook this past month by 600 million plus users
- Youtube
 - **More than 2 billion videos** were watch on YouTube yesterday
- Twitter
 - **32 billion searches** were performed last month on Twitter

Source of Big Data

(1) public data

(2) private data

(3) data exhaust

- Information-seeking behavior, people's needs, desires, or intentions

(4) community data

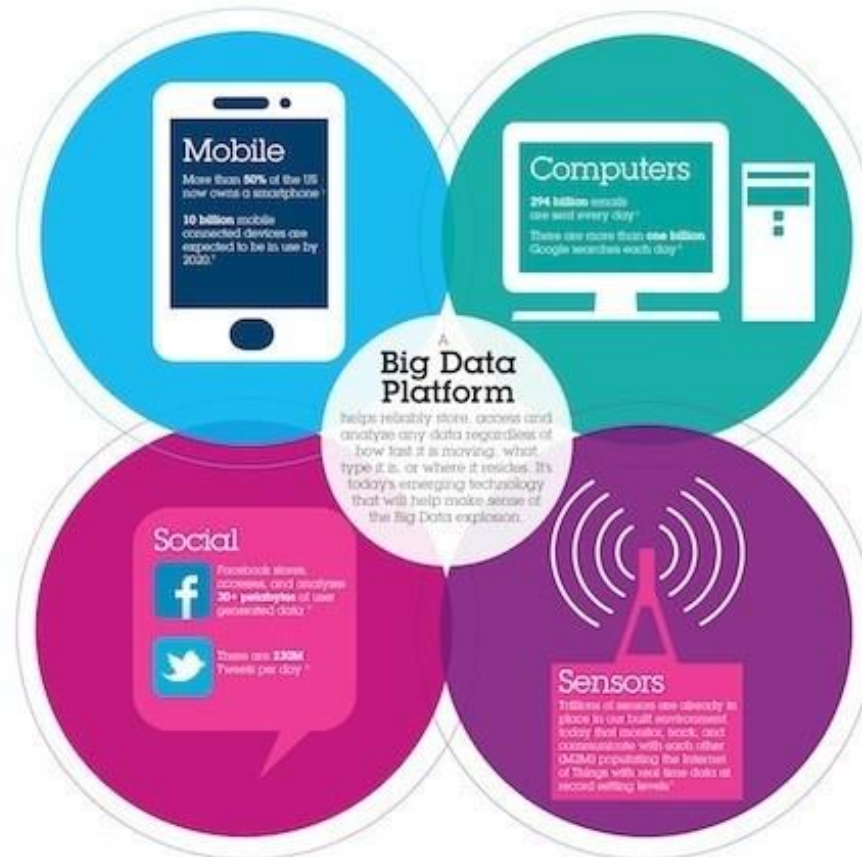
- Customer reviews on products, voting buttons, Tweeters feeds

(5) self-quantification data

- Quantifying personal actions and behaviors, wristbands

Big Data: Making the World go Round

Big Data is growing and moving fast from a variety of sources; are you keeping up?



Information gathered by IBM:

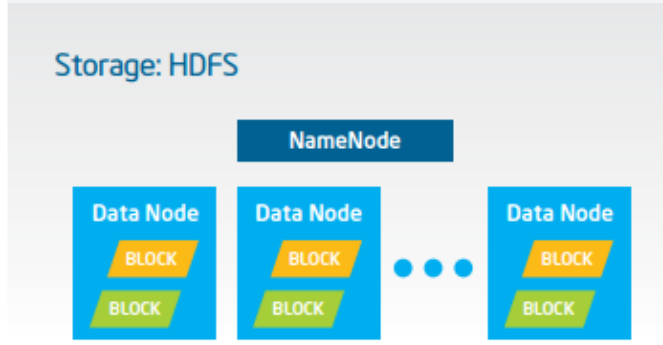
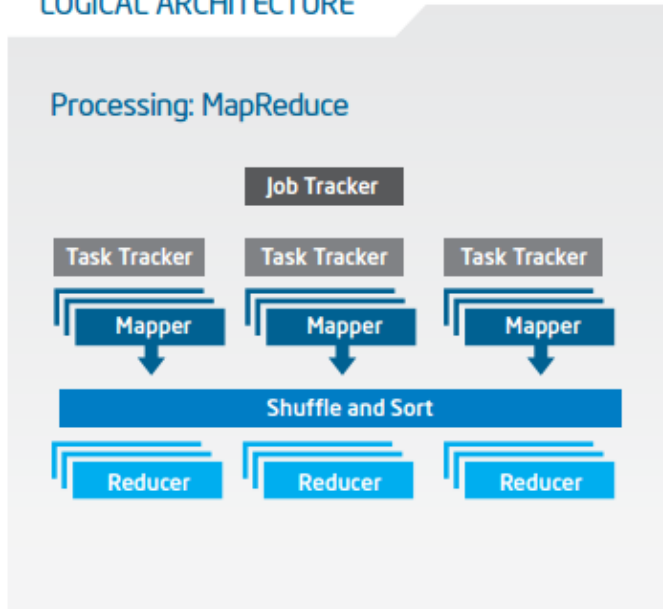
- 1. Clarks Analytics Consulting - US Mobile Data Market Update Q3 2012
- 2. 2011 SensorNet
- 3. IBM - Managing the Big Flood of Big Data in Digital Marketing
- 4. Google - How Google Search Works
- 5. Wikibon - Taming Big Data
- 6. IBM - Managing the Big Flood of Big Data in Digital Marketing
- 7. IBM

IBM, the IBM logo, and the IBM name are trademarks of International Business Machines Corporation. Other product names and service trademarks belong to their respective owners. © IBM Corporation 2012. All rights reserved.

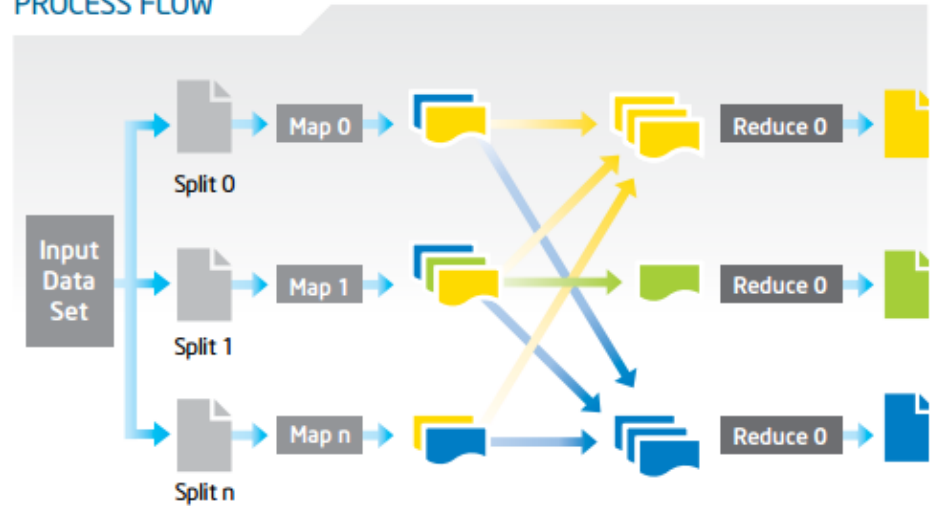


Big Data with Hadoop Architecture

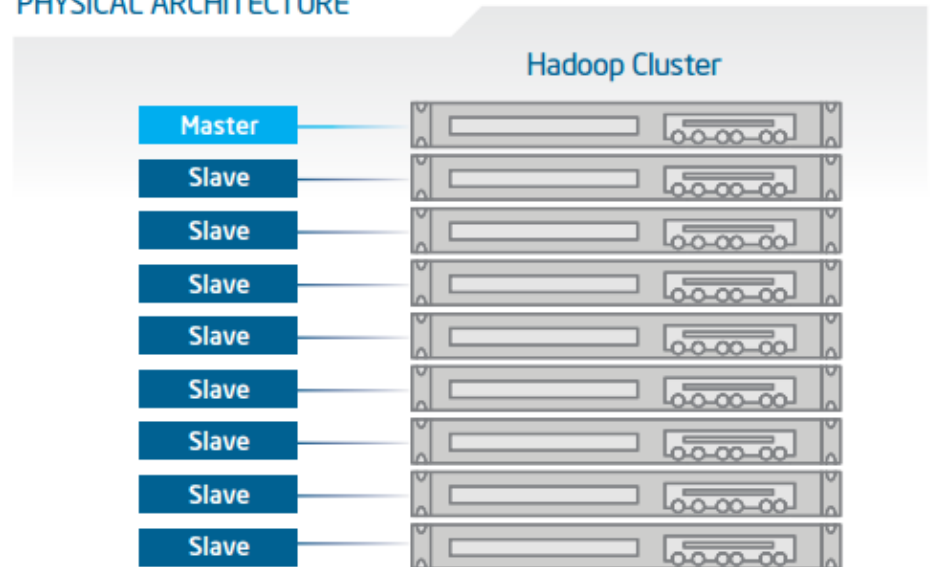
LOGICAL ARCHITECTURE



PROCESS FLOW



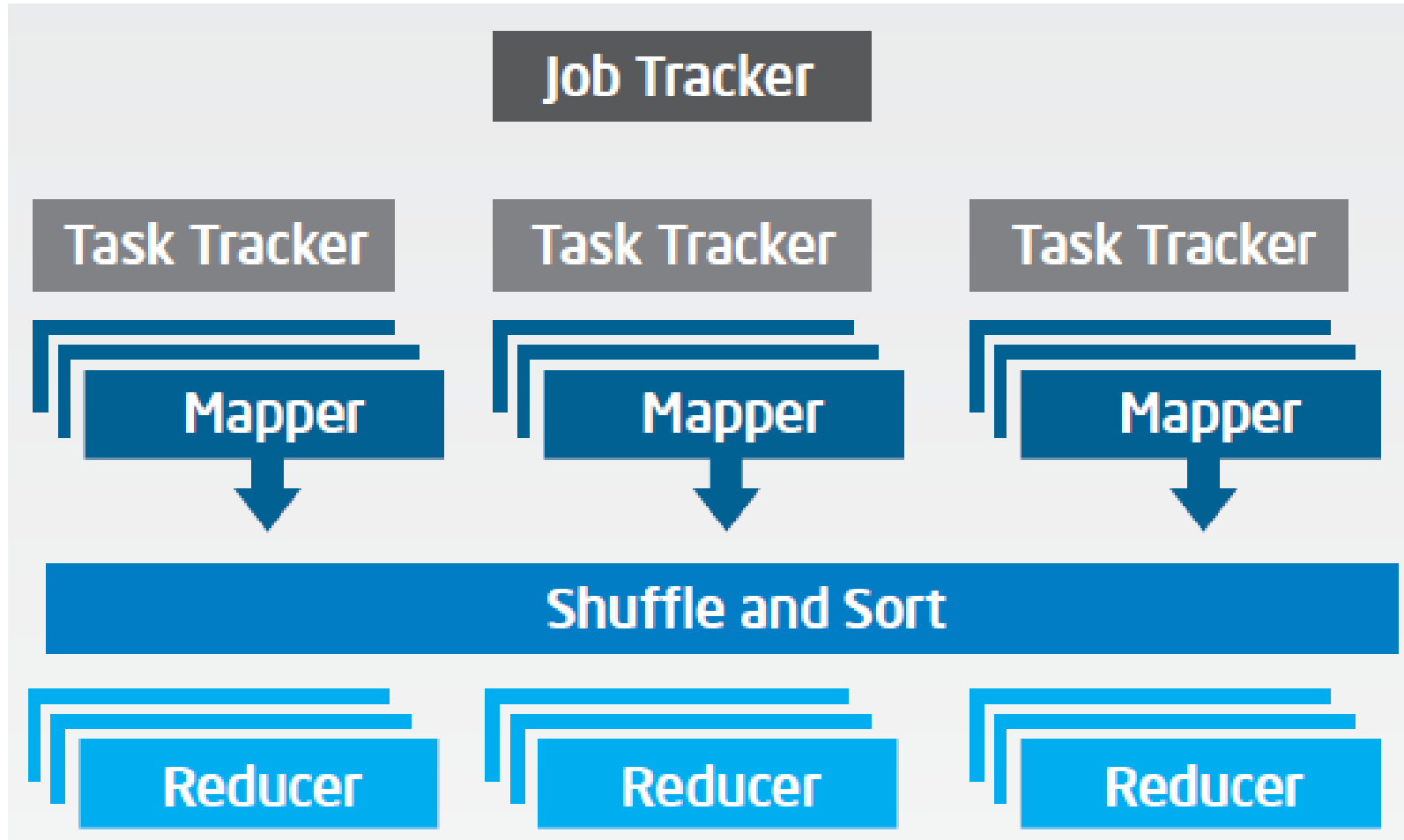
PHYSICAL ARCHITECTURE



Big Data with Hadoop Architecture

Logical Architecture

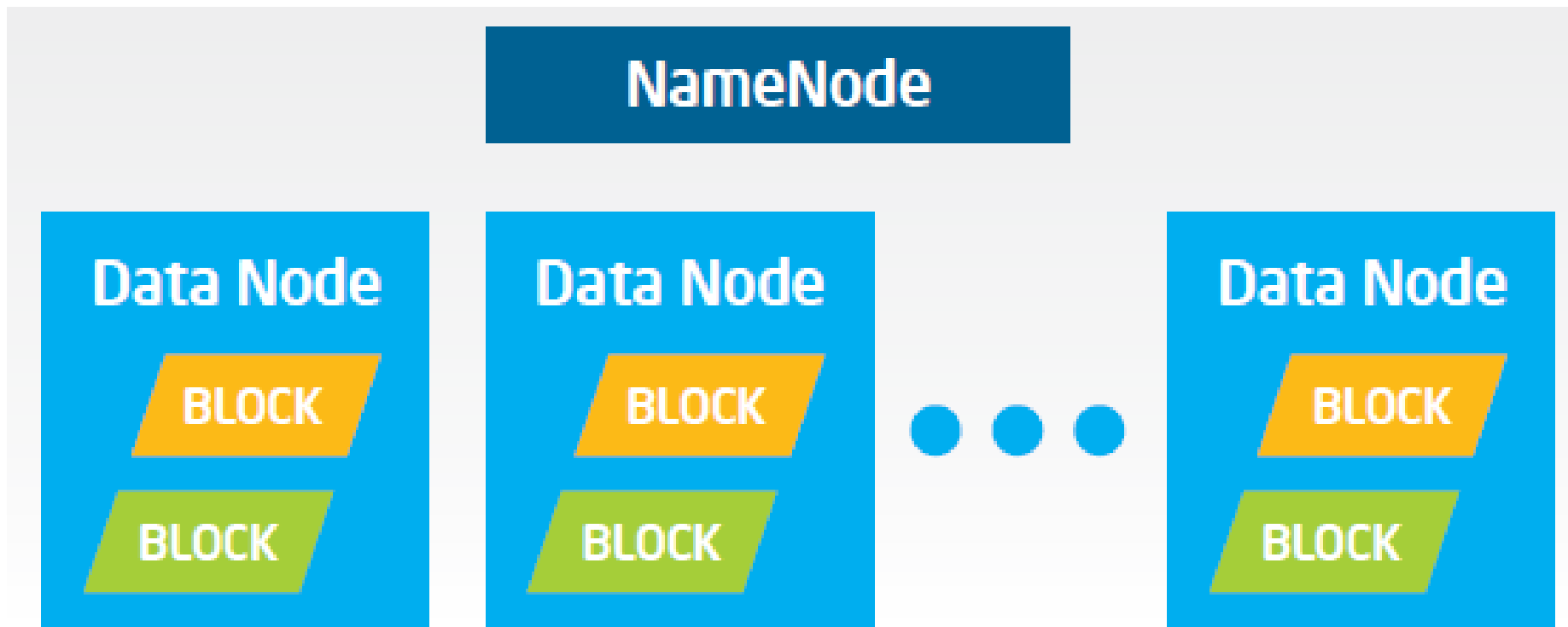
Processing: MapReduce



Big Data with Hadoop Architecture

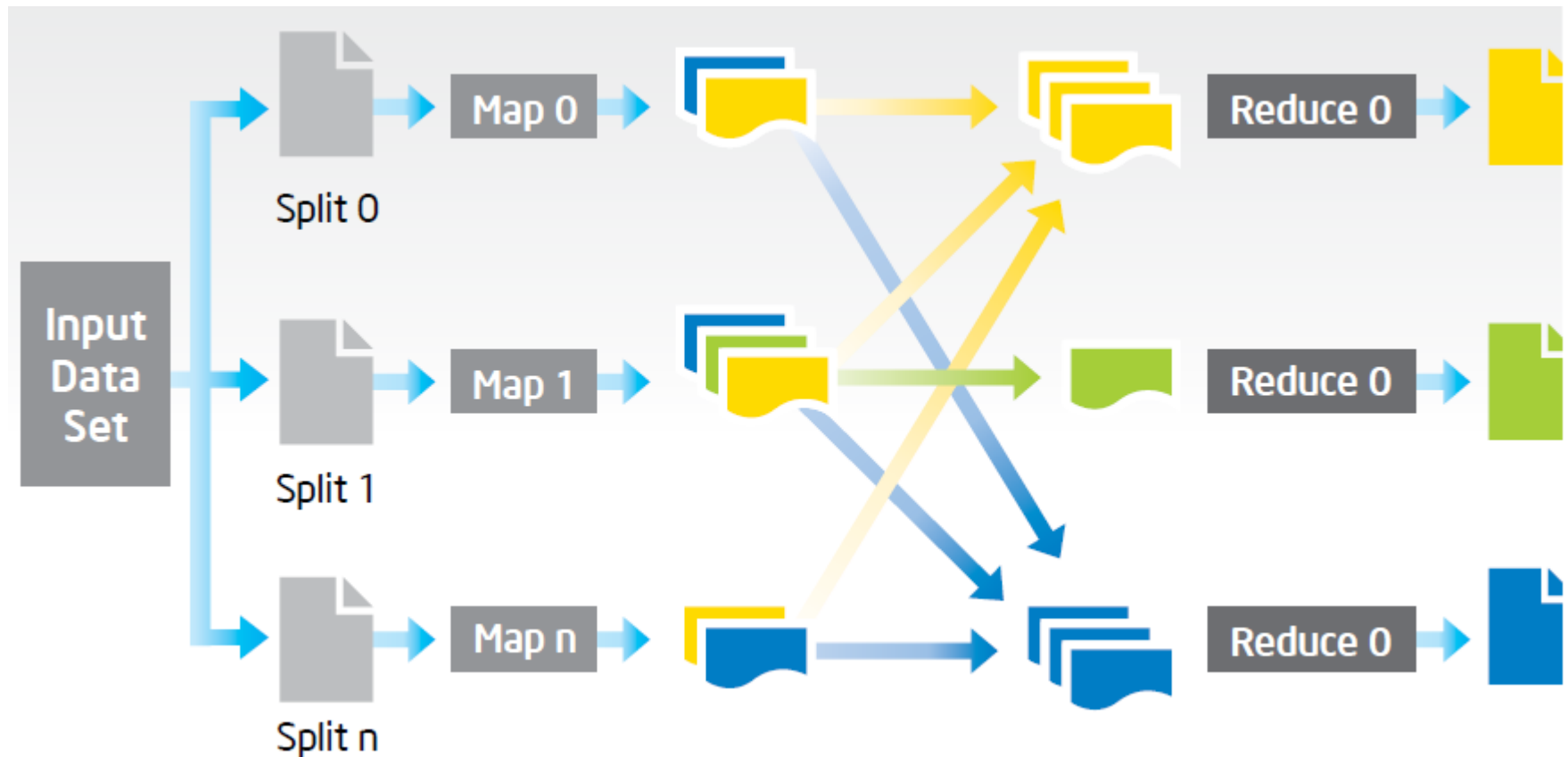
Logical Architecture

Storage: HDFS



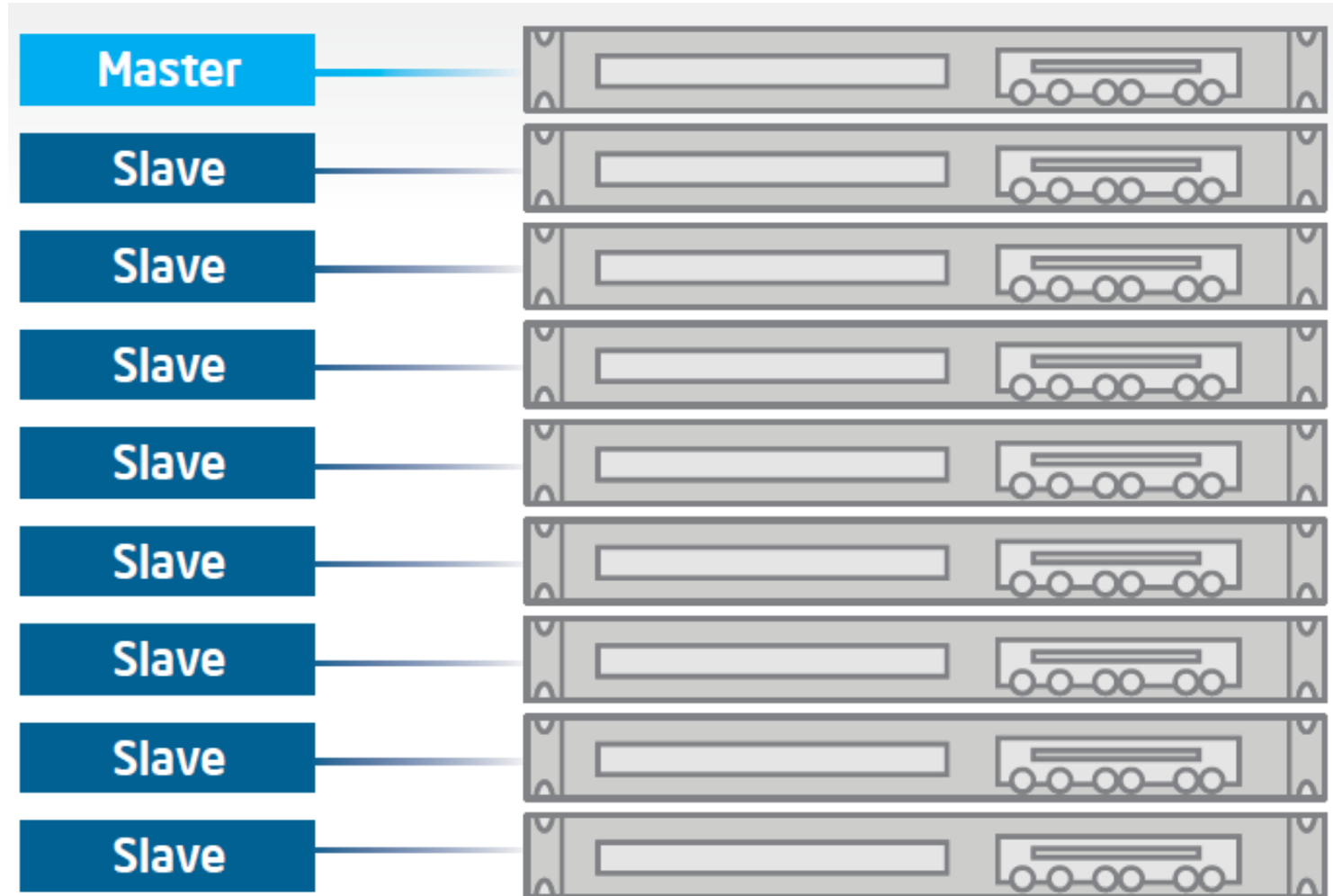
Big Data with Hadoop Architecture

Process Flow

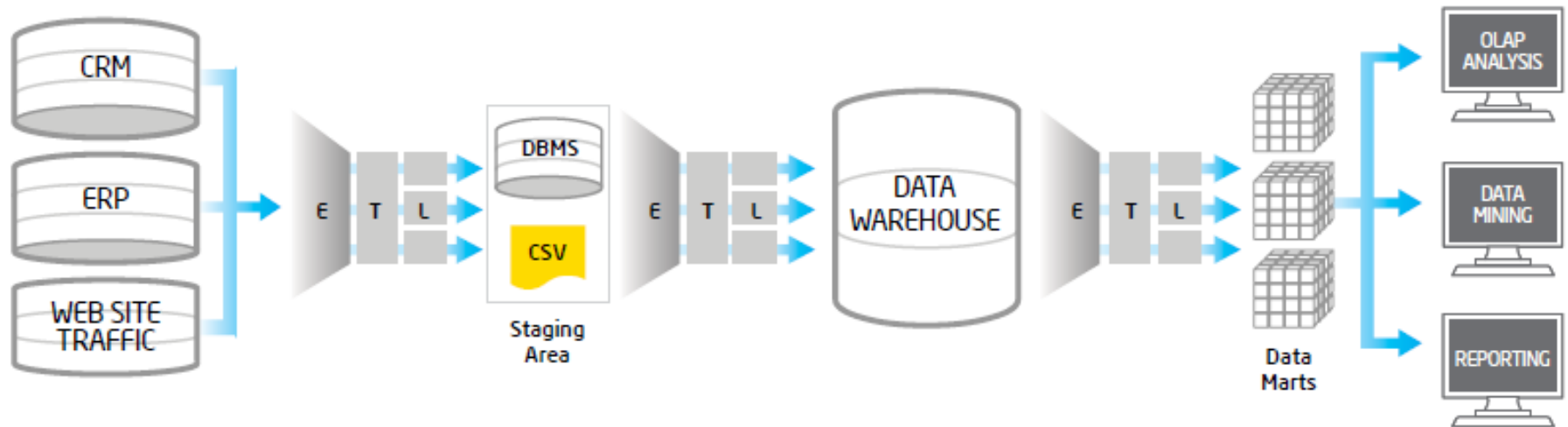


Big Data with Hadoop Architecture

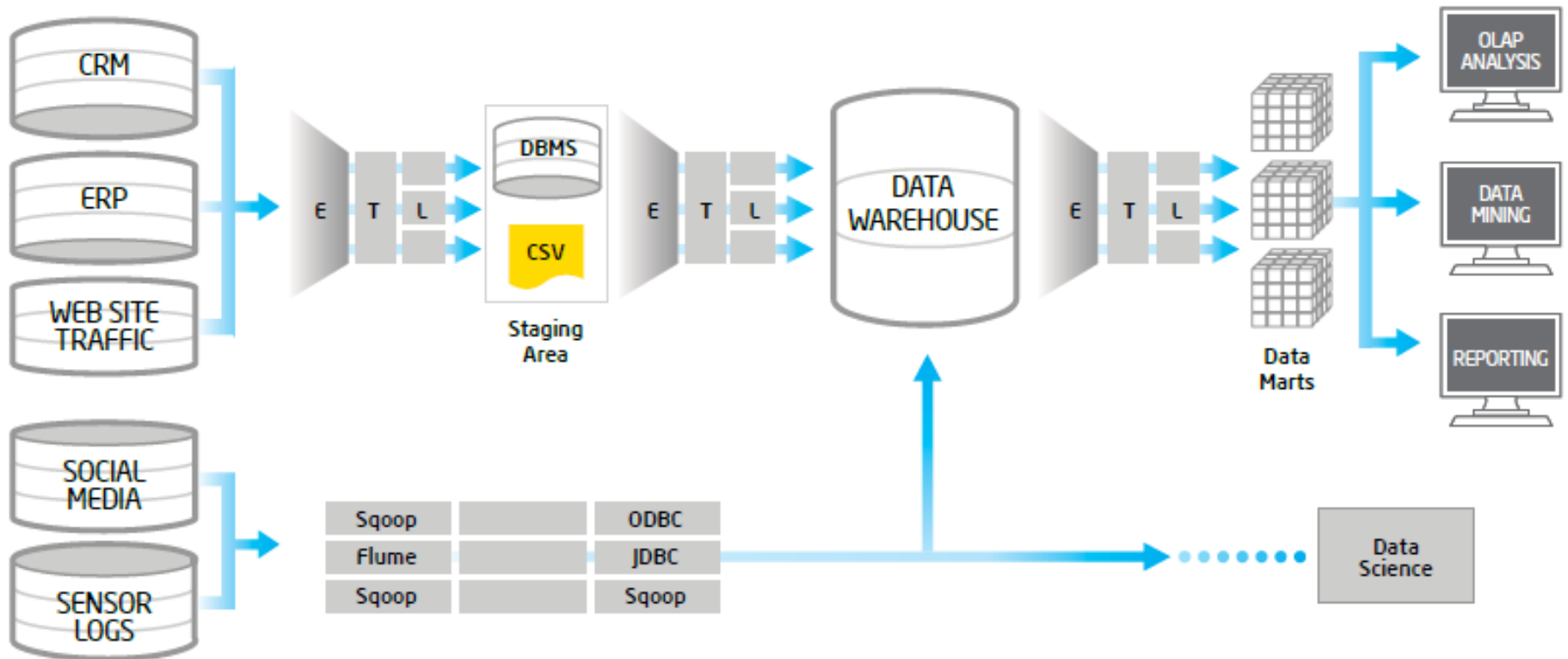
Hadoop Cluster



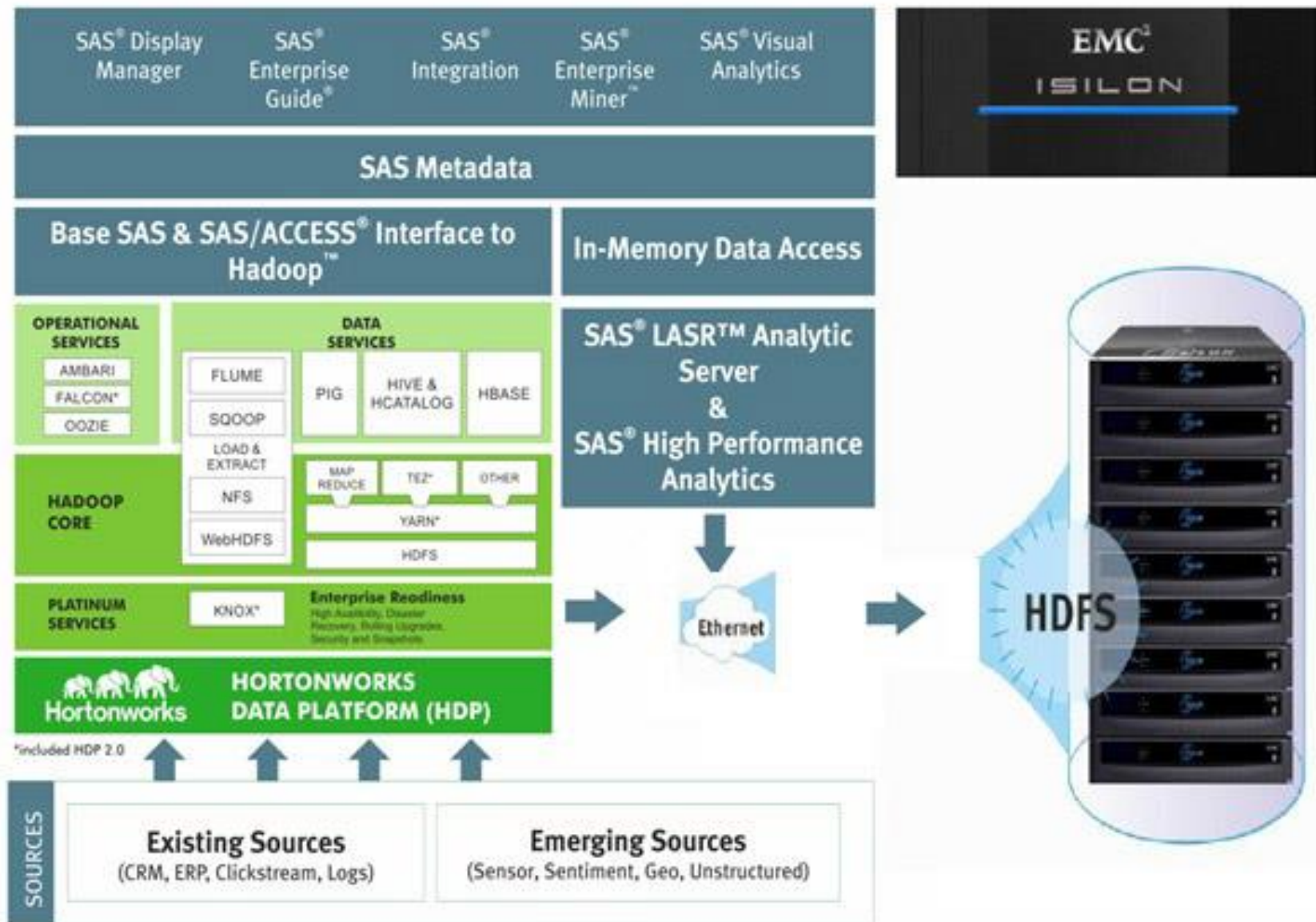
Traditional ETL Architecture



Offload ETL with Hadoop (Big Data Architecture)

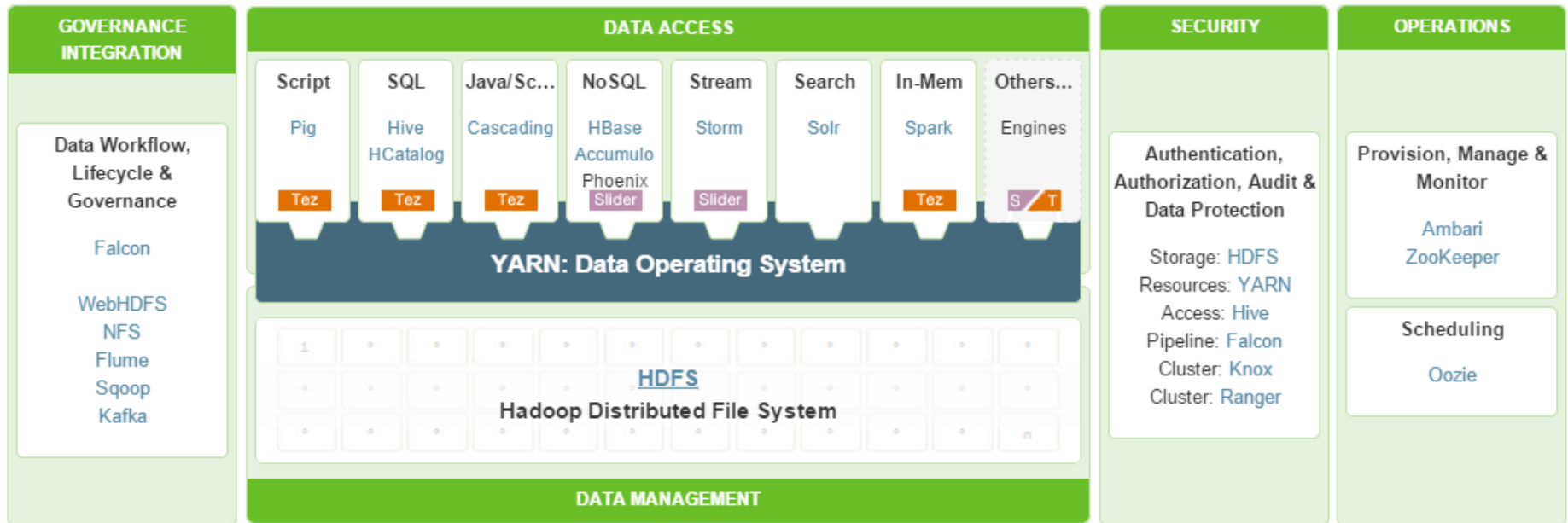


Big Data Solution

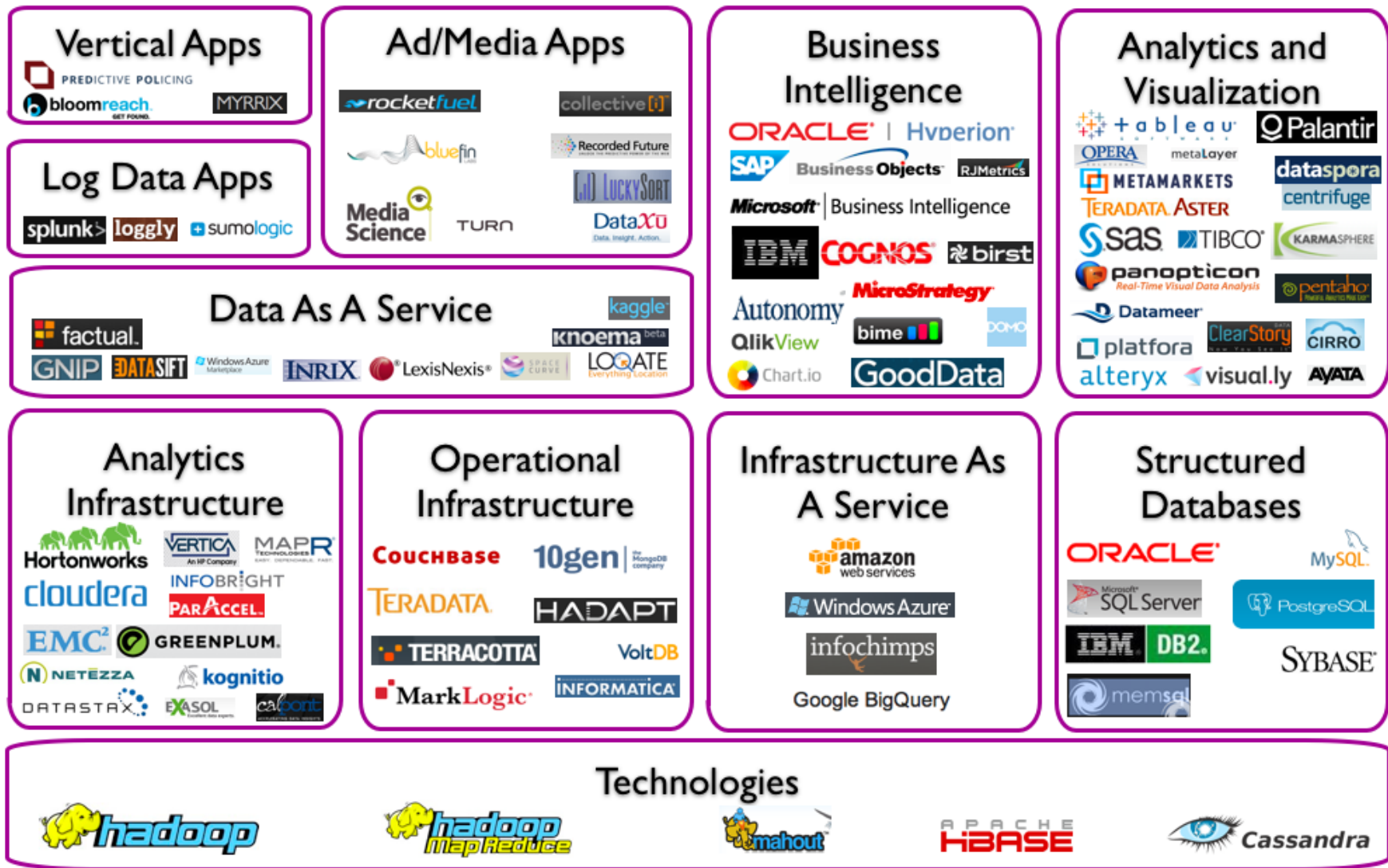


HDP

A Complete Enterprise Hadoop Data Platform



Big Data Landscape



Big Data Landscape (Version 2.0)

Infrastructure



Analytics



Applications



Cross Infrastructure / Analytics



Open Source Projects



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

Source: <http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/>

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

Infrastructure

NoSQL Databases
 FOUNDATIONDB, DATSTAX, mongoDB, COUCHBASE, ZEROSPIKE, HYPERKTABLE, aqri, CLOUDANT, OhnData, Neo4j, sonos

Hadoop On Prem
 HADAPT, cloudera, splice, Zettaset, amazon, MAPR, Microsoft, Hortonworks, Pivotal, IBM

Big Data
 MORTAR, infochimps, Quobole, JETRO, altiscale, AMAZON REDSHIFT

NewSQL Databases
 MarkLogic, TRANSATLICE, Plain, paradigm, memsql, deep db, skySQL, NUODB, Clustrix, VoltDB, SQLFire

Cluster Service
 LexisNexis, HPCC Systems, mesosphere, Acunu

Management / Monitoring
 BUTER, THOUGHT, New Relic, metator, StackIQ, tidemark, appnomic, oceanSYNc, DATADOG, boundy

MPP Databases
 TERADATA, ParStream, InfiniDB, Kognitio, NETEZZA, Pivotal, PARACCEL, SQL Server

Graph Databases
 Neo4j, aster data, InfiniteGraph

Data Transformation
 TRIFACTA, Paxata, KALIDO, Evelytix, syncsoft

Security
 DATAGUISE, Stormpath, IMPERIA

Storage
 Cleversafe, panasas, sinbstrorage, Compuverde

App Dev
 CONTINUITY, wibedata

Crowd-sourcing
 microTASK, servio, mobileworks

Analytics

Analytics Platforms
 databricks, STAT WING, CIRRO, TREPAREL, PERSASIVE, QUAVUS, Datameer, KARMA360, collectiveIQ, FREDDO, dataspora

For Business Analysts
 OrigamiLogic, ClearStory, DataGravity

Data Science Platforms / Tools
 domino, Alpine, Sense, MORTAR, CONTINUUM, ploty, yhat, MODE

Unstructured Data
 BASIS, ATTIVO, GENERAL SENTIMENT, semantria, crimson hexagon, ai Quid, Palantir

Data Visualization
 tableau, ZebraData, visual.ly, Roambi, Chart.io, looker, Ayasdi, ISS, DataHero, TICKBOARD

BI Platforms
 birst, Jaspersoft, pentaho, GoodData, Platfora

Machine Learning
 SKYTREE, big ml, YOUTAINE ANALYTICS, wise.io, contact relevant

Social Analytics
 simple reach, bitly, synthesio, Dataminr, Statistik, DATA SIFT, track, bottlenose

Analytics Services
 THINK BIG, Valance, DATA SCIENCE, MU SIGMA

Statistical Computing
 SAS, MATLAB, Kibana

Log Analysis
 splunk, loggly, sumologic

Location / People / Events
 RADIUS, Fliptop, LOCATE, PlaceIQ

Big Data Search
 hp, LucWorks, ONTOLOGY

Crowd-Sourced
 kaggle, METAMARKETS, DataKind

Real Time
 amato, causita

SMB
 RJMetrics, retention, sumail, GoSquared, custora

Applications

Ad Optimization
 aggregate knowledge, rocketfuel, TAPAD, ai Match, MediaMath, 33across

Publisher Tools
 Chartbeat, Yieldex, yieldbot

Marketing
 LATTICE ENGINES, Sailthru, spinnkr, gainsight, Kontera, RelateIQ, Tell apart, persado, bloomreach, CLICKFOX, Pursuway

Finance
 Lenddo, BILL GUARD, wonga, cignifi, LendUp, KENSHO, OnDeck

Human Capital
 evolv, Centelo, gild

Legal
 JUDICATA, RAVEL, Lex Machina

Government / Regulation
 mark43, enigma, FORTSCALE, feedzai

Security
 SCIFYD, sift science, FORTSCALE, feedzai

Education / Learning
 KNEWTON, eclara, PANORAMA, Clever

Health
 Recombine, 23andMe, Ginger.io, FLATIRON, Counsyl

Industries
 tubular, OPOWER, SIGHT MACHINE, THE CLIMATE CORPORATION, NEXT BIG SOUND

Cross Infrastructure /

SAP, SAS, IBM, Google, Microsoft, vmware, amazon, 1010data, talend, TERADATA, hp, NetApp

Open Source

Framework
 Hadoop, Spark, HDFS

Query / Data Flow
 Cascandra, SciDB, ORACLE, HBASE, mongoDB, riak, Sqoop

Data Access
 HBASE, mongoDB, riak, Sqoop

Coordination / Work-flow
 ZooKeeper, talend

Real Time
 Storm

Stat Tools
 SciPy

Machine Learning
 MLlib

Cloud Deploy
 Heroku

Search
 Solr, LUCENE, elasticsearch

Data Sources

Data Mkts
 Windows Azure Marketplace, blueka, DataMarket, factual, knoema

Data Sources
 DATA GOV, premise, YODLEE, xignite, VALIDIC, plaid, quandt, STANDARD TREASURY, human/api

Sensor Data
 kinsa, SKYCATCH, STREETLINE, fitbit, RunKeeper, JAWBONE, LUMASENSE TECHNOLOGIES, Withings, BASIS, estimate

Incubators & School
 zipfian, GA, INSIGHT, DataLine

Big Data Vendors and Technologies



<p>Data Acquisition</p> <p>Including Complex Event Processing (CEP) tools</p>	<p>VLDW and BI Appliances</p>	<p>Analytics</p>	<p>BPM & Action</p>
<p>Data Providers</p> <p>And all your own data And your partners data</p>	<p>No SQL</p>	<p>Data Virtualization</p> <p>COMPOSITE SOFTWARE</p>	<p>Microsoft</p> <p>Capgemini - Capping IT off Manuel Sevilla - 2012</p>
<p>Data Governance</p>	<p>BI Tools</p>		

Processing Big Data

Google



Source: http://whatsthebigdata.files.wordpress.com/2013/03/google_datacenter.jpg

Processing Big Data, Facebook



References

- Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, 25(3), 443-448.
- George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, 57(2), 321-326.
- Lim, E. P., Chen, H., & Chen, G. (2013). Business Intelligence and Analytics: Research Directions. *ACM Transactions on Management Information Systems (TMIS)*, 3(4), 17
- Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12), 64-73.
- Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1), 3.
- Khan, Z., Anjum, A., Soomro, K., & Muhammad, T. (2015). Towards cloud based big data analytics for smart future cities. *Journal of Cloud Computing: Advances, Systems and Applications*.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*.
- Davenport, T. H., & Patil, D. J. (2012). Data Scientist. *Harvard business review*.
- Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59.
- Yeoh, W., & Koronios, A. (2010). Critical success factors for business intelligence systems. *Journal of computer information systems*, 50(3), 23.
- Thomas H. Davenport, *Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data*, FT Press, 2012
- Michael Minelli, Michele Chambers, Ambiga Dhiraj, *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley, 2013
- Viktor Mayer-Schonberger, Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt, 2013
- Jared Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*, Wiley, 2014