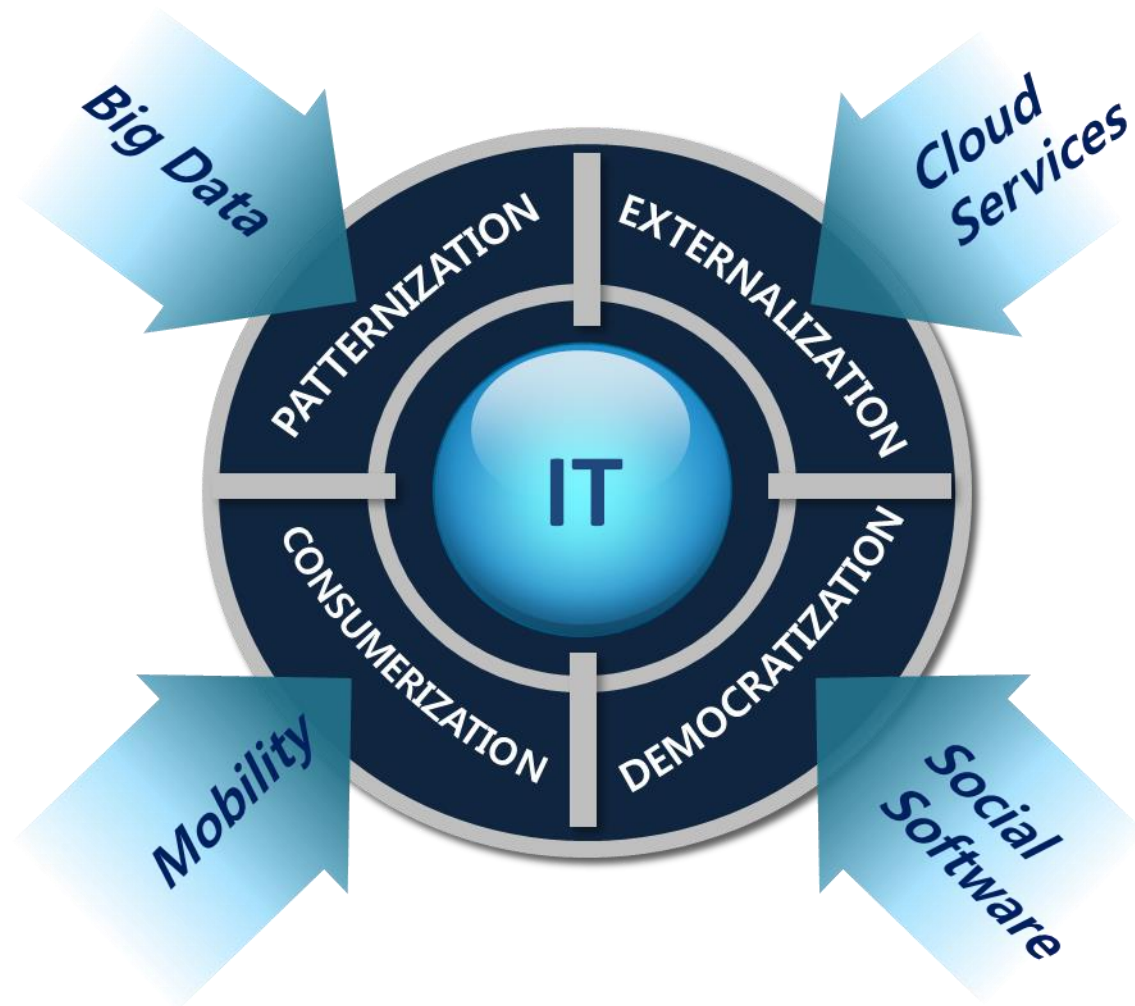


Convergence of Cloud, Mobile, Social and Big Data

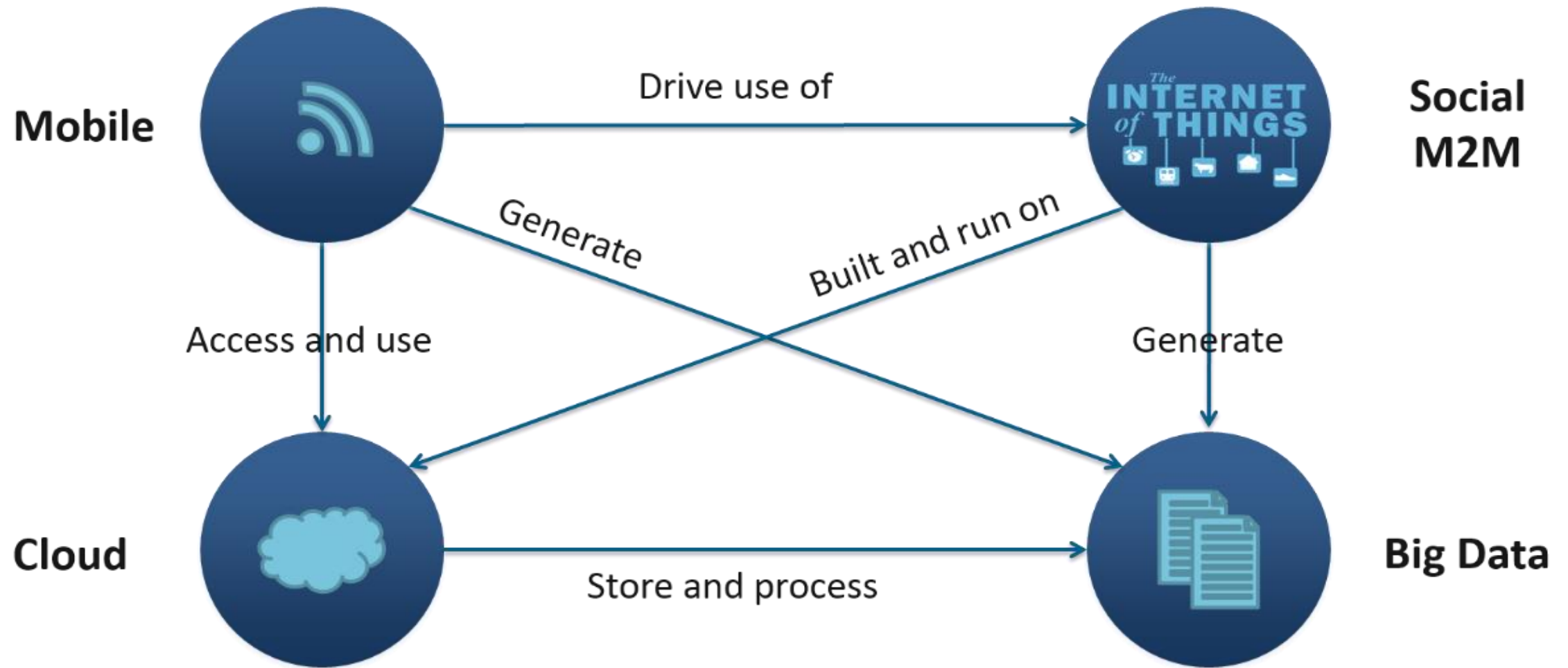
Professor June Sung Park
KAIST

Copyright © 2014. Dr. June Sung Park. All rights reserved.

Convergence of Current IT Trends

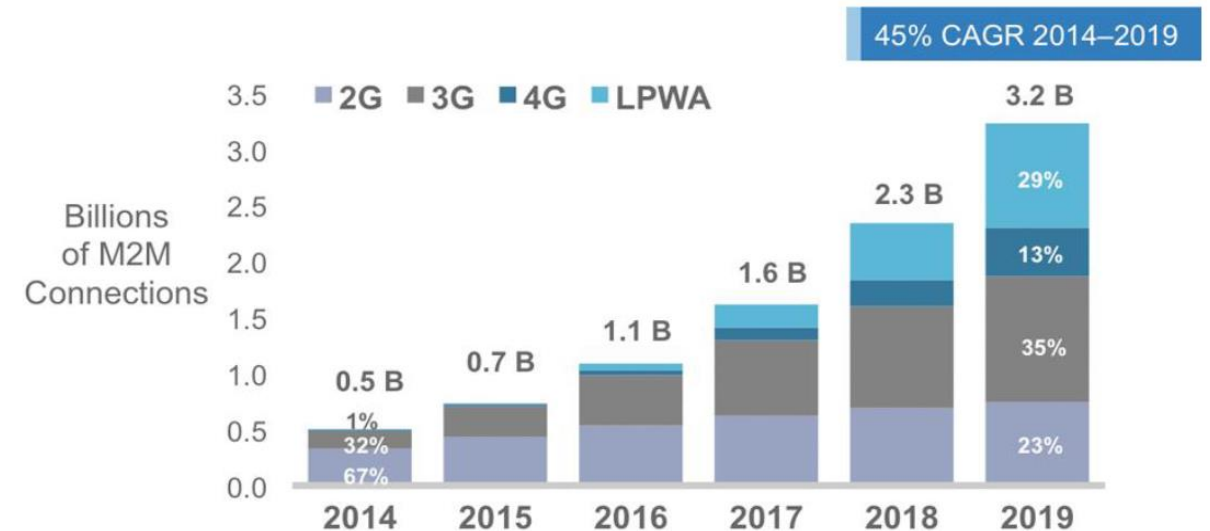
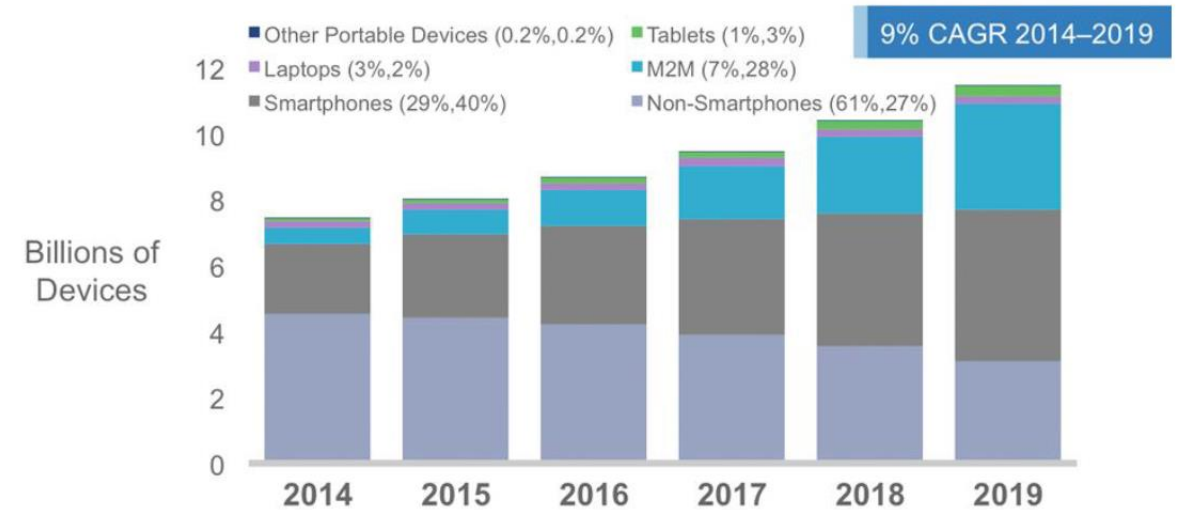


Convergence of Current IT Trends



Mobile Traffic Trend

- Mobile data traffic will grow nearly 10 fold to 24.3 EB per month during 2014- 2019.
- By 2019 there will be 8.9 billion hand-held or personal mobile-ready devices.
- By 2019 there will be 578 million wearable devices, growing 5 fold from 2014.
- By 2019 there will be 3.2 billion M2M connections (e.g., connected car, connected healthcare, asset tracking system).



In 2014, 4G accounts for 1% and LPWA accounts for 0.2% of global mobile M2M connections.
Source: Cisco VNI Mobile, 2015

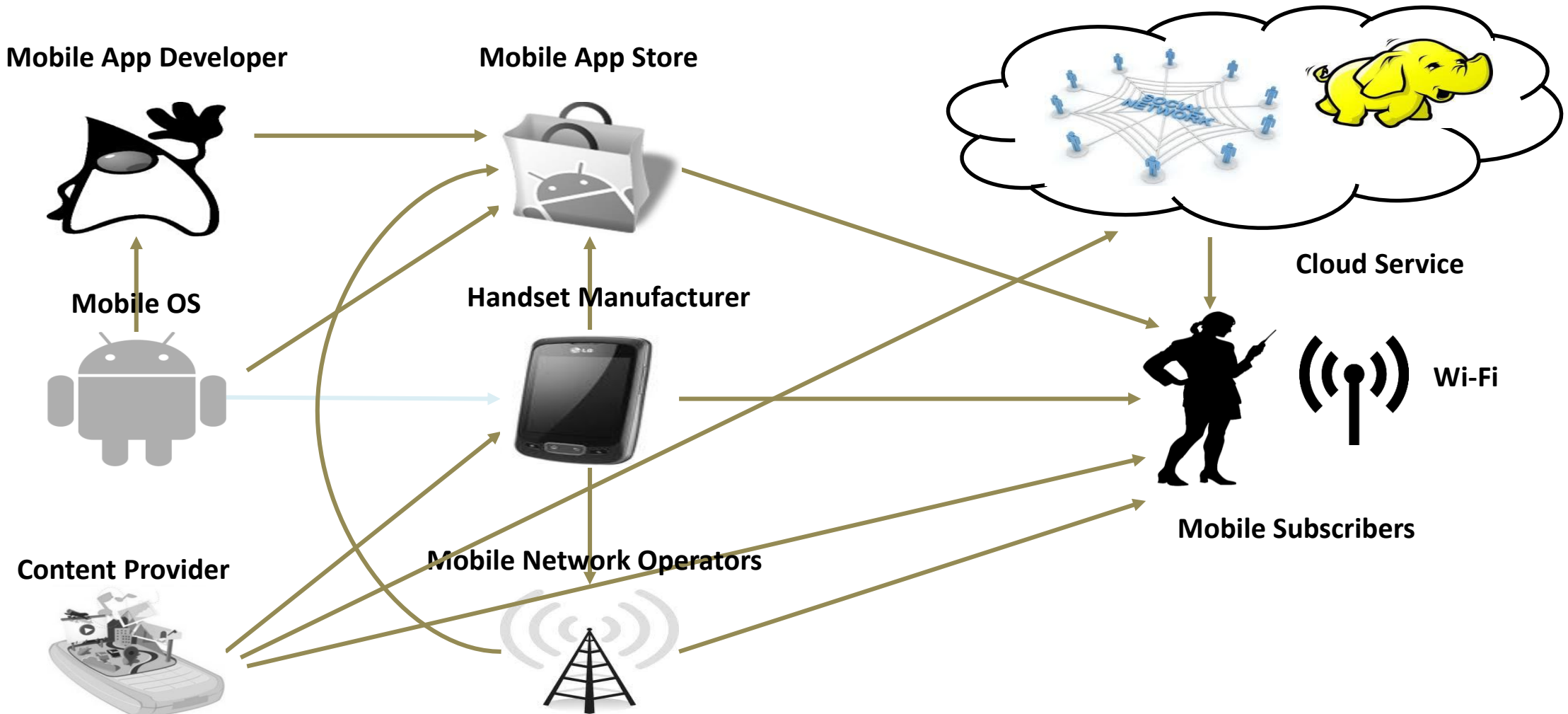
Mobile Cloud

- The market for cross-platform mobile apps on cloud (mobile SaaS) is fast growing at an average annual increase of 88% from annual revenue of \$400 million in 2009 to \$9.5 billion in 2014.
- Mobile cloud traffic will grow 11 fold from 2014 to 2019, accounting for 90% of total mobile data traffic by 2019.
- 75% of the mobile cloud-based app market is represented by enterprise users.
- Dual personas blur professional and personal lines. Business users will demand a unified cloud experience—same apps and contents from home PC, office PC and smart phone on the road.



Mobile Cloud

- Cloud-based mobile apps have the power of a server-based computing infrastructure accessible through an app's mobile Web interface.



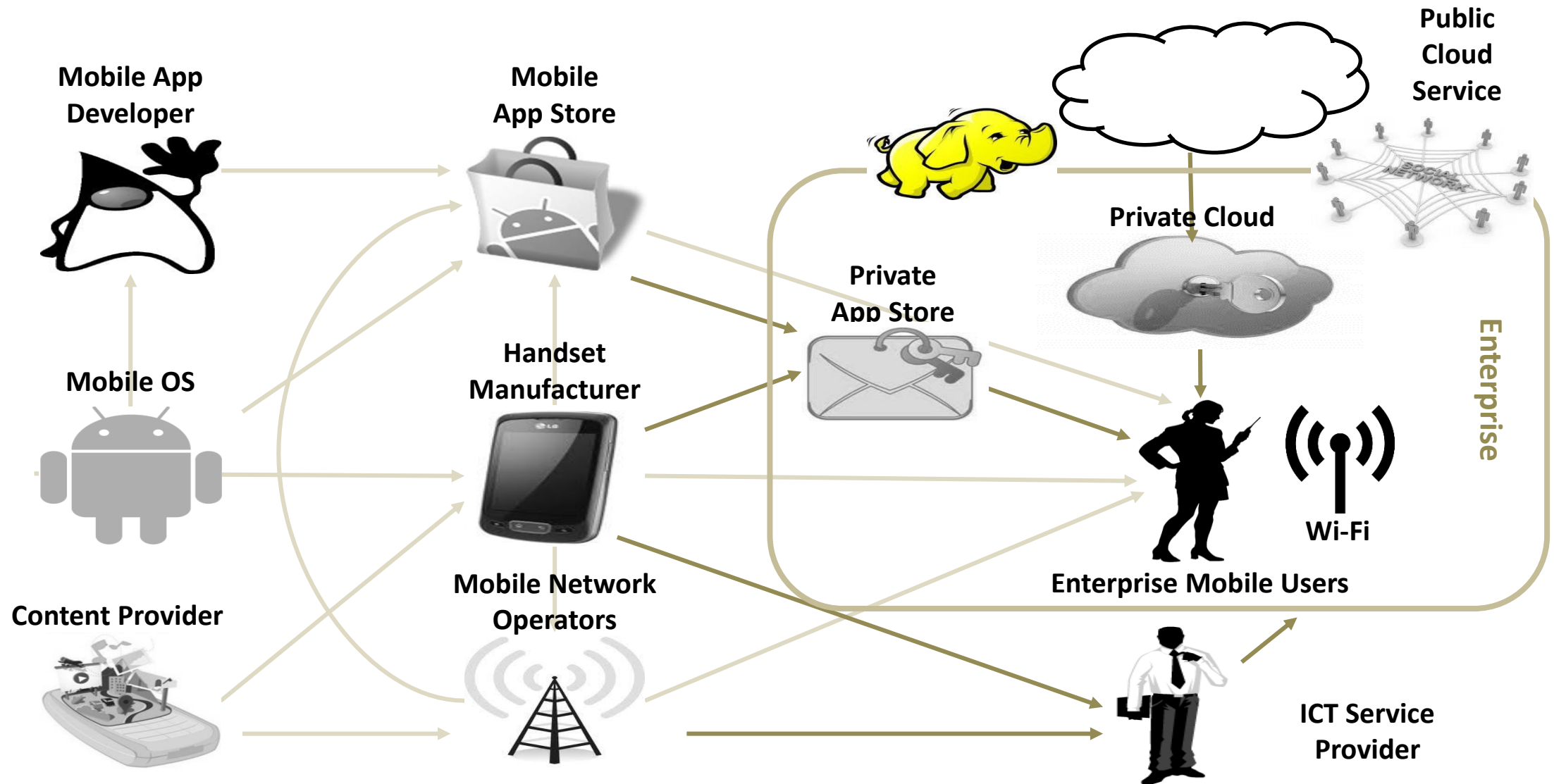
Enterprise Mobility (Mobile Business)

- Three of the top 5 new technology investment priorities from the 2015 Gartner CIO Agenda Survey are what we call “nexus forces”: BI/analytics, cloud and mobile.

| Rank | Investment priority | 2014 | 2015 |
|------|----------------------------------|------|------|
| 1 | BI/analytics | 41% | 50% |
| 2 | Infrastructure and data center | 31% | 37% |
| 3 | Cloud | 27% | 32% |
| 4 | ERP | 26% | 34% |
| 5 | Mobile | 24% | 36% |
| 6 | Digitalization/digital marketing | 17% | 11% |
| 7 | Security | 13% | 11% |
| 8 | Networking, voice and data comms | 12% | 12% |
| 9 | Customer relationship/experience | 11% | 8% |
| 10 | Industry-specific applications | 9% | 10% |
| 11 | Legacy modernization | 7% | 7% |
| 12 | Enterprise applications | 6% | 2% |

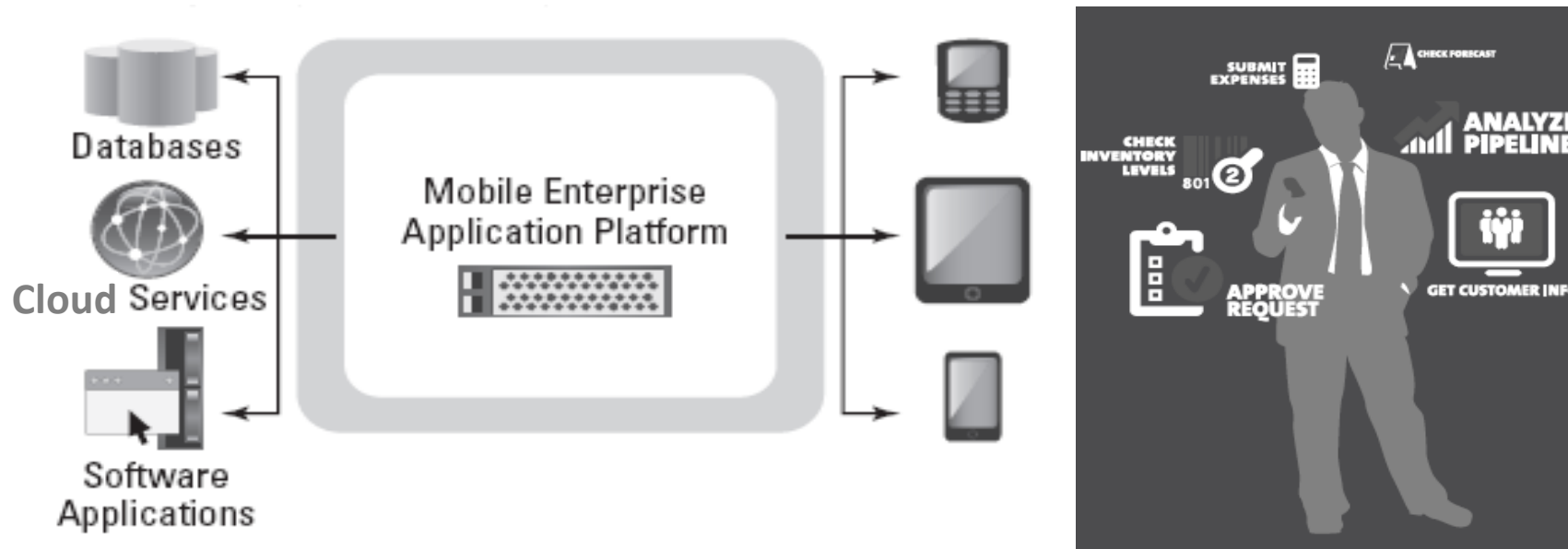
n = 2,793

Enterprise Mobile Cloud



Enterprise Mobile Cloud

- Mobile Enterprise Application Platform (previously called Multiple-Channel Access Gateway) provides an integrated environment for development, execution and management of mobile applications.
- MEAP allows companies to develop an mobile application once and deploy it to a variety of mobile devices, mobile OS, networks and user groups.
- Mobile Backend as a Service (MBaaS) started in 2011 provides MEAP functionalities from the cloud, bridging mobile frontends with cloud-based backends via unified API and SDK.



Use Cases of Enterprise Mobile Cloud



On-Demand Mobile Services

- Hospitals such as New York-Presbyterian Hospital, Mayo Clinic, Kaiser Permanente have developed a medical information platform that provides patient data to both physicians and patients.

Companies developing new online, on-demand, mobile services as SaaS for sales, logistics, healthcare, etc.



Mobile Streaming Services

- Walt Disney and Warner Bros both are developing cloud-based systems to provide consumers with instant access to films and TV shows via cable, computers and smart phones.

Companies developing new on-demand, streaming mobile services for entertainment, surveillance, m-learning, etc.

Use Cases of Enterprise Mobile Cloud



Context-Aware Services

- Ford car software generates data on its location, speed, braking and wiper use. It then correlates the data with live information from the Web about traffic and weather, and sends messages about road conditions via Twitter to other motorists in the same area.

Manufacturers embedding cloud-enabled services inside their products to provide context-aware mobile services.



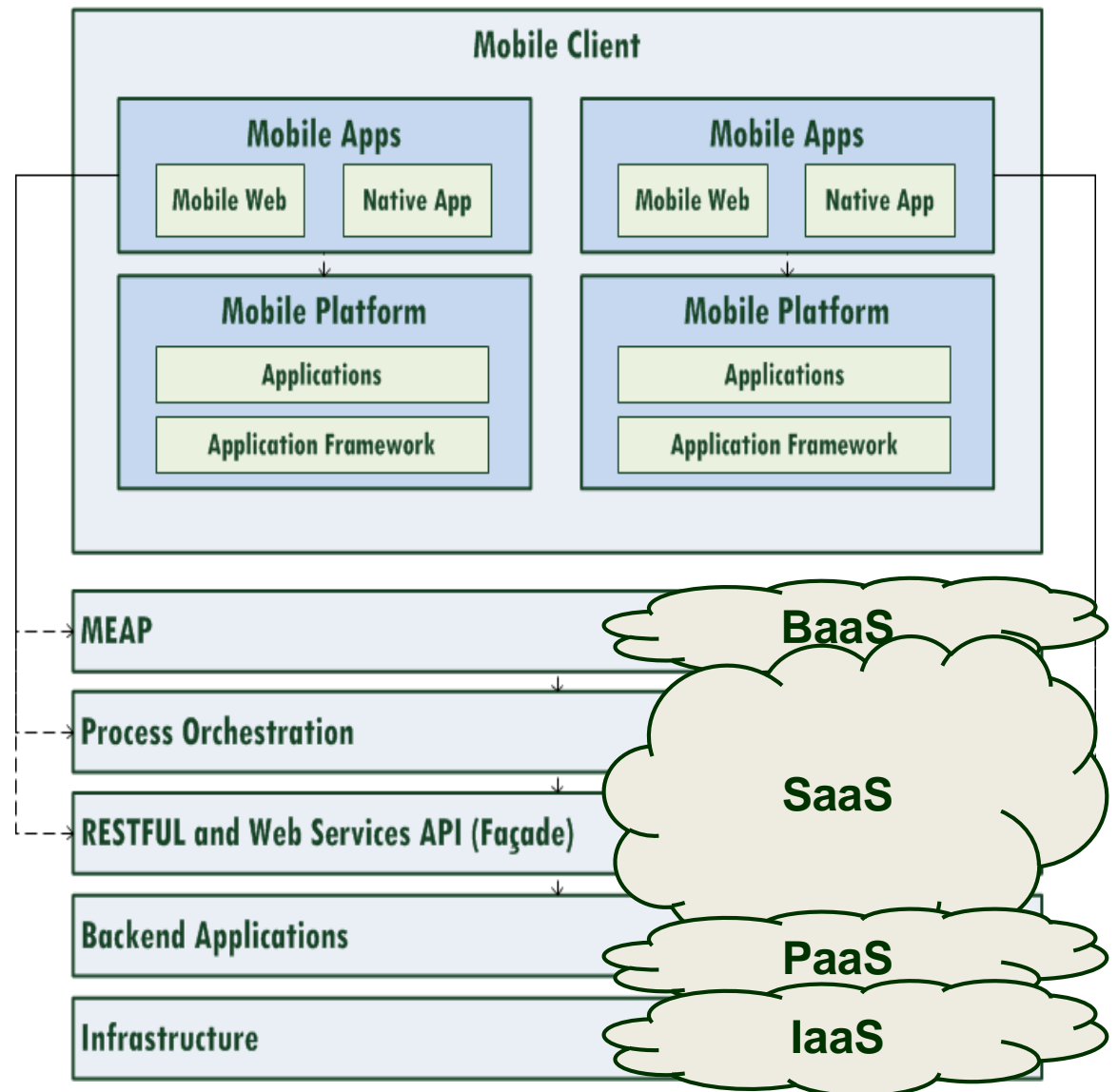
Social Sensing

- Social Sensing is the activity performed by users who act as monitors and provide information needed for an enterprise or its customers to adapt in real time to context changes.

Companies distribute mobile devices with special sensors to crowd-source real-time data they need.

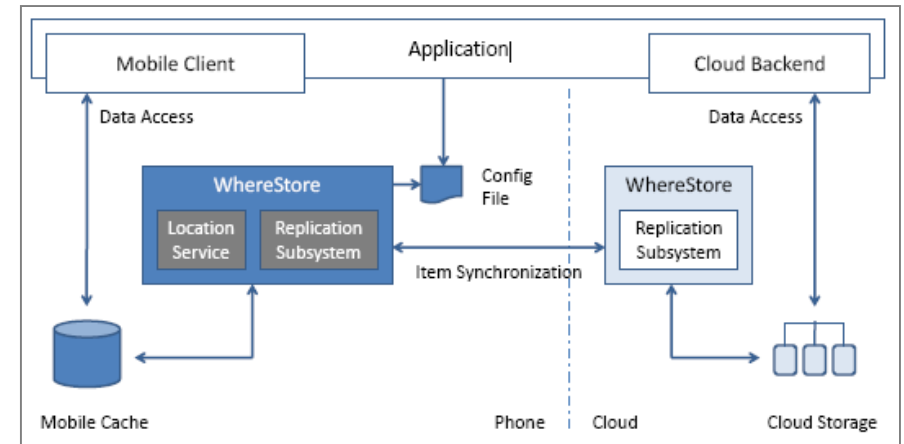
Mobile Cloud Architecture

- Mobile cloud services must be immediate, visual and simple.
 - Shift of control from touch to voice and motion (e.g. Siri, Kinect)
 - Heads-up display (e.g. Google Glasses)
- Mobile clouds require omnichannel delivery.
 - Mobile clouds require a mix of native code and Web view to blend capabilities of native platforms with the cross-platform portability of Web.
 - Modularized services APIs come in handy as user interface technology shifts.
 - Netflix to smart phones, tablets, laptops, TVs, game consoles
- Mobile clouds require dynamic composition of RESTful services.
 - Services remixed from Google, Facebook, Twitter, Twilio, Spotify,... everyday



Mobile Cloud Architecture

- Mobile clouds require use of open source SaaS and PaaS .
 - OpenMEAP now available on OpenShift, and AWS Marketplace
- Mobile clouds require elastic IaaS.
 - Instagram adding 1M new customers in 12 hours
- Mobile clouds must overcome last-mile connectivity bottleneck of 3G/4G through data and computation offloading.
 - Microsoft's WhereStore prototype allows caching data to mobile devices from clouds to reduce data unavailability and access latency.
 - KAIST Smart Cloudlet Research Center develops P2P cloudlet services on D2D network (e.g. Wi-Fi Direct)
 - In-memory databases inside mobile devices (e.g.. Memcached, Ehcache)
- Mobile clouds can provide the roles of remote control, routing, and data/computation offloading for wearables and connectables.
 - Nike+ FuelBand tracks your daily activities and calorie burned.



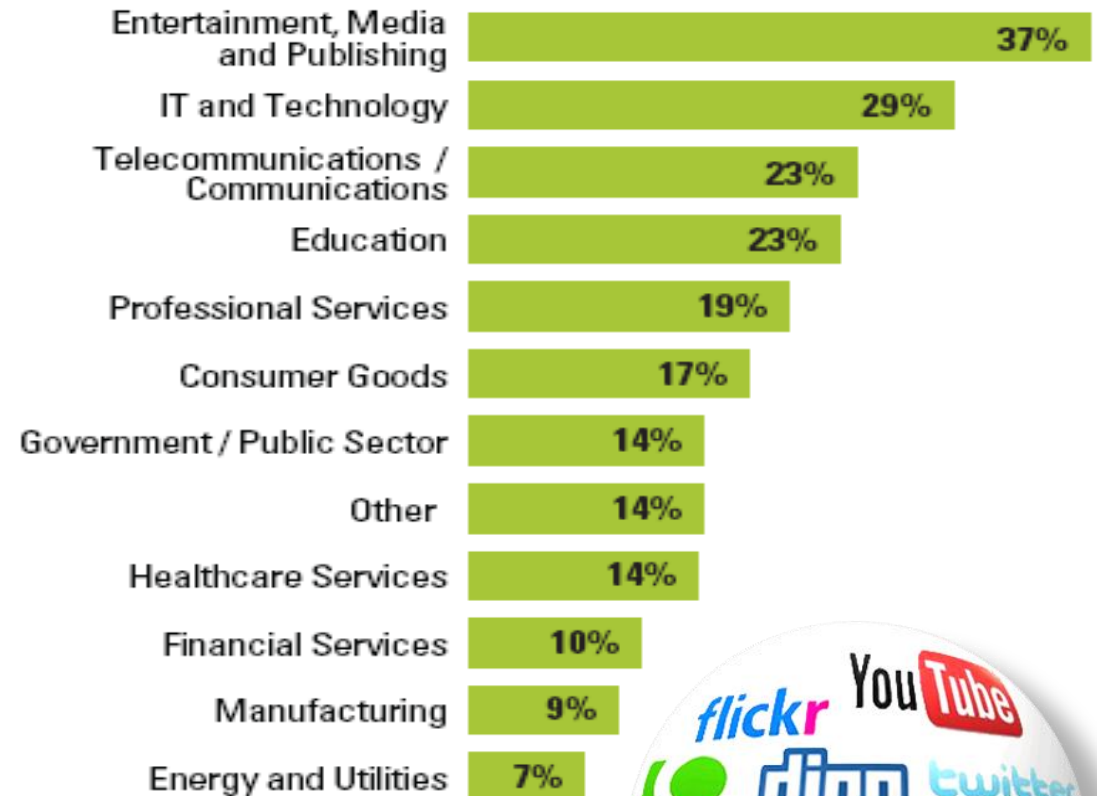
Context-Aware Mobile Cloud

- Mobile clouds can be context aware.
 - Google MyTracks records your path, speed, distance and elevation while you walk, run or bike.
- User's mobile context is recognized via sensors.
 - GPS, NFC, gyroscopes, barometer, accelerometers, microbolometer, magnetometer, chemical sensor,...
- User's mobile context is:
 - Situational: current time, location, altitude, environmental conditions, travel speeds,...
 - Preferential: historical personal decisions
 - Attitudinal: feelings and emotions implied by actions and logistics
- User's mobile context is stateful.
 - Manage states in clients (subject to device capacities) so that you can allow atomic services, stateless, asynchronous communications with clouds and load balancing and linear scale-out for IaaS
- Mobile clouds can provide predictive analytics.
 - Use clouds to store and process the exploding amount of mobile contextual data (Big Data), and to run predictive analytics.
 - Google Now attempts to get you just the right information at just the right time by predicting your actions.

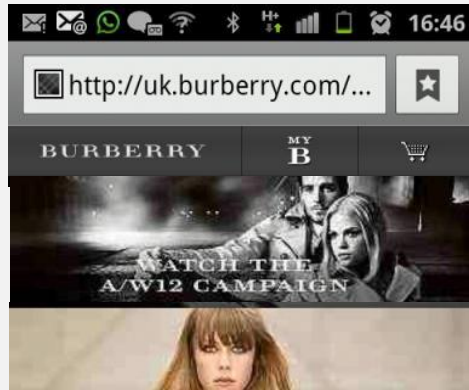


Social Cloud and Social Business

- Social business is defined as business activities that use social software (which is a SaaS).
- The importance of social business to enterprises is growing, and the line between real business and social business is diminishing.
- Entertainment, media and publishing industries and IT/technology industries tend to see the most value from social business today.
- Marketing and innovation are top uses of social business so far.



Use Cases of Social Business



Social Marketing

Social activities help catalyze e-commerce via Facebook, Twitter, YouTube, ...

- Personalized and context-sensitive interaction, conversation and engagement
- Social sharing and recommendations among stakeholders



Social Innovation

- Innovation funnel that allows discovering and embedding new solutions and expertise more rapidly than an internal R&D lab might accomplish.

Use Cases of Social Business



Social Operation

- Whether your company spans seven cubicles or seven continents, the enterprise social network (ESN) connects you to the people, information and content you need to get work done.
- Social technologies are extremely helpful in terms of scaling participation in knowledge flows.



Social Analytics

- Social analytics scours and analyzes a huge variety of sources of content such as news, blogs, tweets and other online media across the Internet.
- Marriott Hotel scans the public social cloud for customer feedback and if a customer posts a complaint, the local hotel is alerted to address the problem.

Data Deluge

- 🌐 All Shakespeare writings are 50 megabytes (10^6).
- 🌐 30 billion pieces of content are shared monthly on Facebook.
- 🌐 A GE jet engine produces 20 terabytes (10^{12}) of data in every operating hour.
- 🌐 From the dawn of civilization until 2003, humankind generated 5 exabytes (10^{18}) of data. Now we produce 5 exabytes every two days, and the pace is accelerating.
- 🌐 Unstructured data such as documents, spreadsheets, email, Web log, images and videos (High Definition and 3D) make up the lion's share of existing data in most enterprises and most of the growth.
- 🌐 Enterprises today process more than 60 terabytes of information annually, about 1,000 times more than a decade ago.
- 🌐 Every animate and inanimate object on earth will soon be generating data, including our homes, our cars, and our bodies.
- 🌐 Big data analytics enables decision makers in an enterprise to quickly spot patterns that affect its competitive advantage.

Big Data

- 🌐 Structured and unstructured, static and streaming data of large volumes on the order of petabytes (10^{15}) which conventional database and data warehouse technologies cannot efficiently store and process
- 🌐 Data accumulated in databases and data warehouses
- 🌐 Data generated from Web (as Digital Footprint; a.k.a. Data Exhaust), sensors and machines



Limitations of Current RDB and ROLAP

- 🌐 Data deluge entailed the increase in volume and variety of data and the required velocity of processing.
- 🌐 Current RDBMS and data warehouse technologies have limited capabilities in handling extreme volumes and varieties of data at a desired velocity.
 - 🔍 A RDBMS was originally designed to run on a single machine. Therefore, scaling with RDBMS typically requires running on a bigger machine (i.e., a scale-up).
- 🌐 Google and Amazon developed different approaches based on distributed parallel processing technologies that allow unlimited scale-outs using low-cost commodity hardware.
 - 🔍 A disk drive that can store all the music in the world (a few TBs) is only \$600.
 - 🔍 However, it takes an average of 2.5 hours to read 1TB.
- 🌐 A new set of data models have also emerged over the last few years (e.g. key-value, document, column family, graph data models, etc.) that are collectively called NoSQL (Not Only SQL):
 - 🔍 NoSQL scale for huge datasets using large clusters of low-cost machines.
 - 🔍 NoSQL require no fixed schema to allow any unstructured data to be stored.
 - 🔍 NoSQL are efficient in storing and processing sparse data.

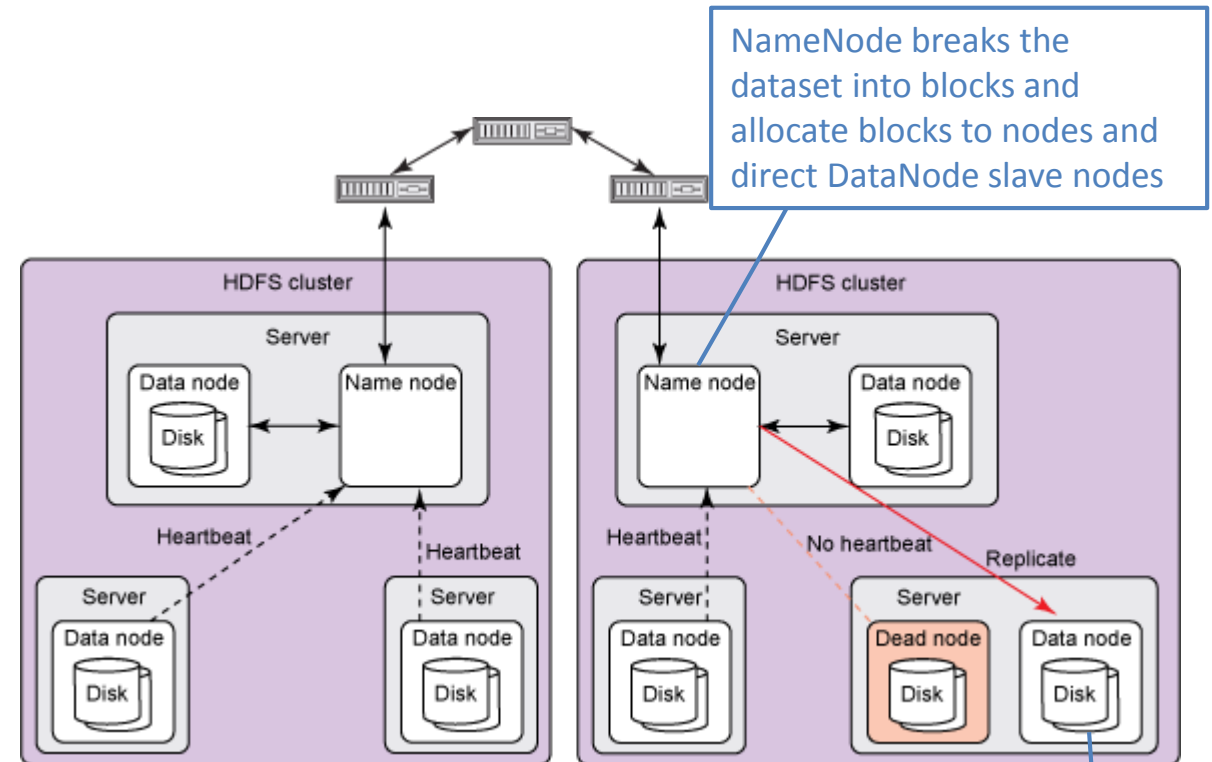
Big Data Platform

- Implement new, specialized analytical platforms designed to accelerate query performance when running complex functions against large volumes of data (with a better total cost of ownership as little as \$10,000 per terabyte).
- Use massively parallel processing (MPP) columnar databases running on commodity servers and distributed file systems running MapReduce and other non-SQL types of data processing languages.
- Hadoop is now central to the big data strategies of enterprises.

| Functional layers | Hadoop subprojects |
|---|--------------------------|
| Hadoop modeling and development | MapReduce, Pig, Mahout |
| Hadoop storage and data management | HDFS, HBase, Cassandra |
| Hadoop data warehousing, summarization, and query | Hive, Sqoop |
| Hadoop data collection, aggregation, and analysis | Chukwa, Flume |
| Hadoop metadata, table, and schema management | HCatalog |
| Hadoop cluster management, job scheduling, and workflow | Zookeeper, Oozie, Ambari |
| Hadoop data serialization | Avro |

Big Data Platform: HDFS

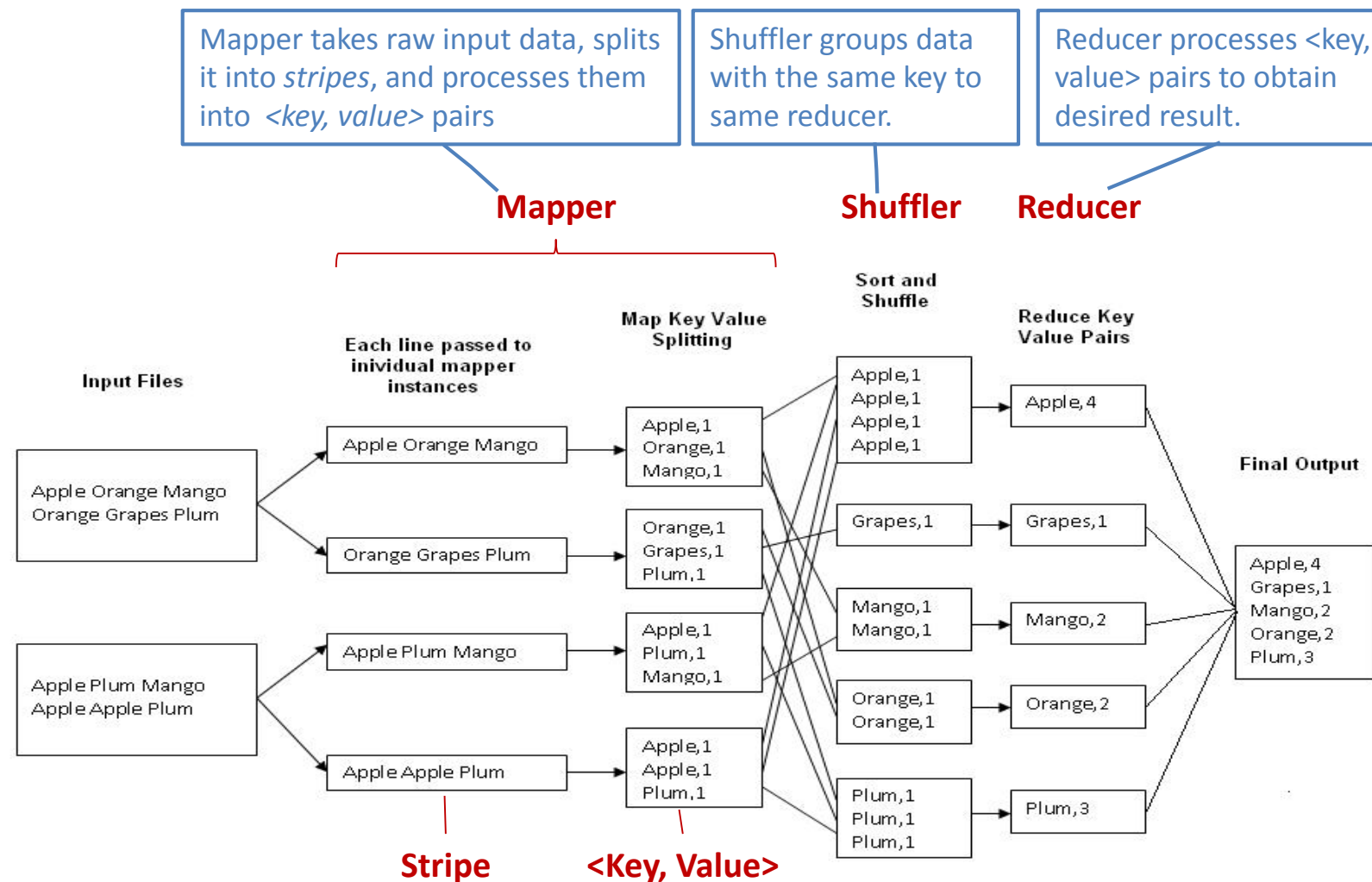
- 🌐 Hadoop Distributed File System (HDFS) created by Doug Cutting at Yahoo in 2005 based on Google DFS, now became top-level Apache project (hadoop.apache.org)
- 🌐 Scale out with low-cost commodity hardware
 - 🔍 Allows storage of large data files > 100 TB (= 25 X 4TB disks)
 - 🔍 If you distribute 1TB on 100 disk drives, it takes 1.5 minutes to read.
- 🌐 Runs on top of native file system on each node
 - 🔍 Input data is split into *blocks* ((usually 64MB or 128MB) each of which is stored as a file on a node.
 - 🔍 A large dataset is stored in many blocks over many nodes.
 - 🔍 Blocks are replicated for reliability.



```
hadoop fs -ls
hadoop fs -mkdir /user/junepark
hadoop fs -put mydata.txt /user/junepark
hadoop fs -cat mydata.txt
hadoop fs -get mydata.txt
```

Big Data Platform: MapReduce

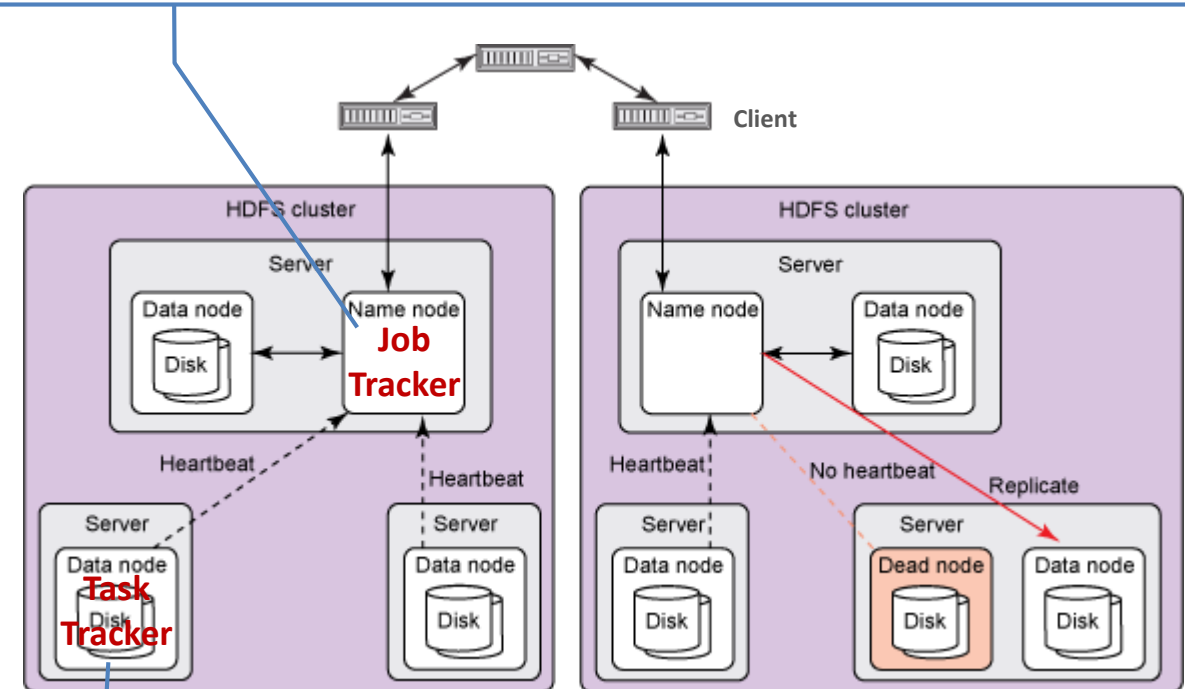
- How can we process a dataset distributed on 100 disk drives and retrieve the result fast?
- Hadoop MapReduce is used as the distributed programming model:
- Map → Shuffle → Reduce.



Big Data Platform: MapReduce

- 🌐 Hadoop MapReduce creates Job Tracker and Task Tracker which collaborate to process MapReduce jobs.
- 🌐 Running a MapReduce job requires:
 - 🔍 Create a Java class(map) that extends Mapper with such parameters as input key(line number), input value(line text), output key (fruit)and output value (count).
 - 🔍 Create a Java class(reduce) that extends Reducer that receives <key, value> pairs with the same key (<fruit, 1> pairs) and returns <key, value> pairs(<fruit, count> pairs).
 - 🔍 Run the Hadoop job by creating a Hadoop Job object, specifying input path for datasets to be processed, Mapper class type, Reducer class type, and output path for the result to be written.
 - 🔍 Once complete, job is submitted to Job Tracker.

Job Tracker receives MapReduce jobs from clients, determine files to process, assign nodes to different processing tasks, and monitors tasks as they execute



Task Tracker manages individual map or reduce tasks that Job Tracker issues, and updates Job Tracker with task progress using heartbeat signal

NoSQL: Pig

- Pig simplifies Hadoop programming.
- Pig provides Pig Latin query language.
- Pig Latin commands can be run in Grunt interactive shell, script files or embedded in Java programs.

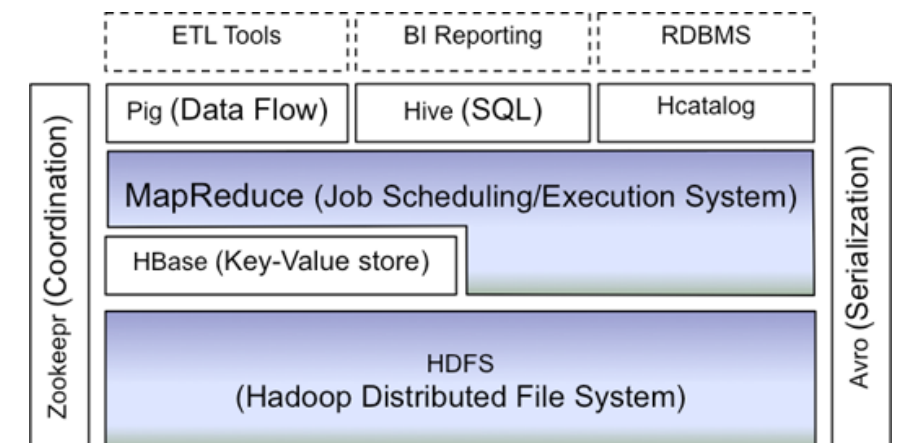
> pig -x local

```
grunt> recordings = LOAD 'recordings.txt' USING PigStorage(',')
>      AS (artist : chararray, title : chararray, duration : double,
>      year : long, price : double);
grunt> artists = GROUP recordings BY artist;
grunt> result = FOREACH artists GENERATE recordings.artist,
>      COUNT(recordings);
grunt> STORE result INTO 'results';
grunt> DUMP results;
```

recordings File in CSV format:

(Artist, Title, Duration, Year of Sale, Price)

```
"Rolling Stones", "Beggars Banquet", , 9.99
"Rolling Stones", "Dear Doctor", ,
"Rolling Stones", "Factory Girl", ,
, "The Killers", ,
"Julia Stone", "By The Horns", 3.45, 2012, 9.99
"Joni Mitchell", "Both Sides Now", 5.46, 2012, 6.95
"Julia Stone", "By The Horns", 3.45, 2012, 7.89
"Lana Del Ray", "Born To Die", 4.46, 2012, 7.99
```



NoSQL: MongoDB

- 🌐 MongoDB organizes data in the Document model which stores of semi-structured text or binary information such as XML, YAML, JSON, BSON, PDF, MS Word, etc.

```
use junepark
album = { id: 1,
          artist: "Rolling Stones",
          title: "Beggars Banquet",
          price: 9.99}
```

```
db.recordings.insert(album)    Insert a document into a collection
```

```
db.recordings.find({artist: "Rolling Stones"})
db.recordings.find({price: {"$gte":5.00. "$lte": 7.00}})
```

recordings File in CSV format:

(Artist, Title, Duration, Year of Sale, Price)

"Rolling Stones", "Beggars Banquet", , 9.99

"Rolling Stones", "Dear Doctor", ,

"Rolling Stones", "Factory Girl", ,

, "The Killers", ,

"Julia Stone", "By The Horns", 3.45, 2012, 9.99

"Joni Mitchell", "Both Sides Now", 5.46, 2012, 6.95

"Julia Stone", "By The Horns", 3.45, 2012, 7.89

"Lana Del Ray", "Born To Die", 4.46, 2012, 7.99

NoSQL: Cassandra

- Column Family data model stores data in tabular format, but without formal schema.
- Each row has a unique key and can have multiple columns.
- Cassandra provides CQL query language.

```
CREATE KEYSPACE junepark;  
USE junepark;
```

```
CREATE COLUMN FAMILY recordings (  
  Title varchar PRIMARY KEY  
  Artist varchar,  
  Price double);
```

**column family schema that
can be changed dynamically**

```
INSERT INTO recordings (Title, Artist, Price) values  
(‘Beggars Banquet’, ‘Rolling Stones’, 9.99);
```

```
SELECT * FROM recordings;
```

recordings File in CSV format:

(Artist, Title, Duration, Year of Sale, Price)

“Rolling Stones”, “Beggars Banquet”, , 9.99

“Rolling Stones”, “Dear Doctor”, ,

“Rolling Stones”, “Factory Girl”, ,

, “The Killers”, ,

“Julia Stone”, “By The Horns”, 3.45, 2012, 9.99

“Joni Mitchell”, “Both Sides Now”, 5.46, 2012, 6.95

“Julia Stone”, “By The Horns”, 3.45, 2012, 7.89

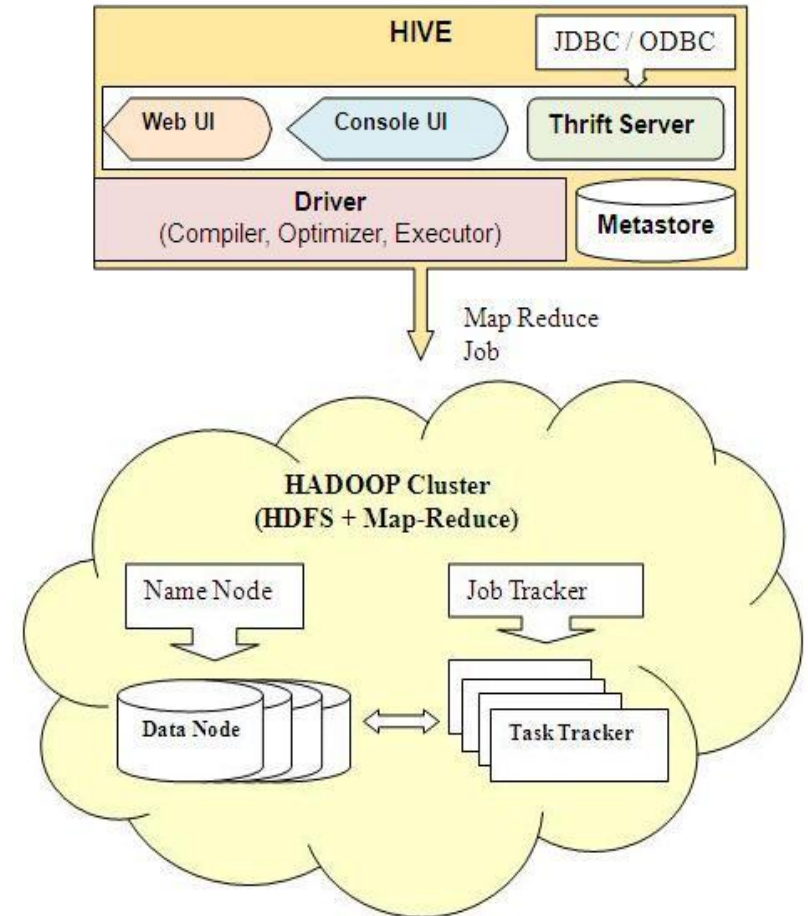
“Lana Del Ray”, “Born To Die”, 4.46, 2012, 7.99

NoSQL: Hive

- Hive is a data warehousing package built on Hadoop by Facebook to handle structured data.
- It provides SQL-like language HiveQL to write ad hoc queries and analyses.
- No need to know about Hadoop programming
- Hive provides HiveQL query language.

```
hive> CREATE TABLE music_recordings (artist STRING, title STRING,  
> duration DOUBLE, year INT, price DOUBLE)  
> ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
> STORED AS TEXTFILE;  
hive> SHOW TABLES;
```

```
hive> LOAD DATA LOCAL INPATH 'data/music_data.txt'  
> OVERWRITE INTO TABLE music_recordings;  
hive> CREAT TABLE recording_frequency (artist STRING, count INT);  
hive> INSERT OVERWRITE TABLE recording_frequency  
> SEKECT artist, COUNT(artist) FROM music_recordings GROUP BY artist;  
hive> SELECT * FROM recording_frequency;
```

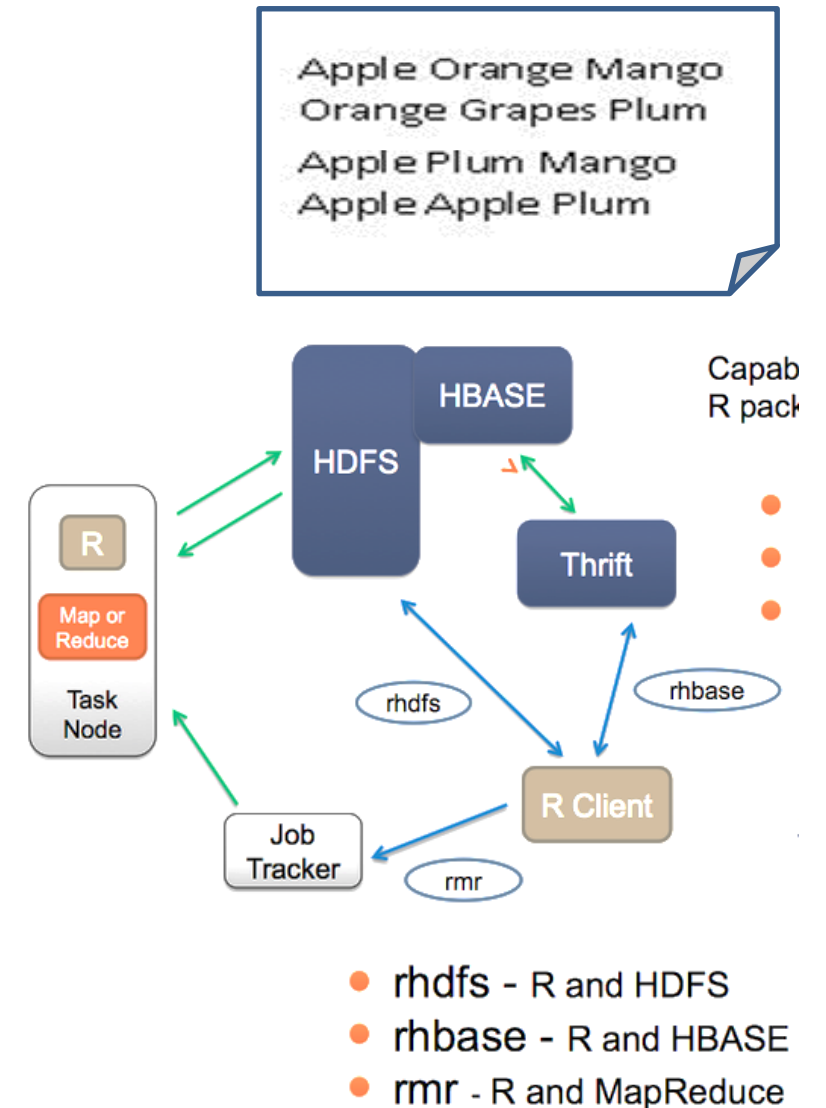


Big Data Analytics: RHadoop

- R is a statistical programming language used for statistical and predictive analytics, data mining and visualization
- Rhadoop, an R wrapper around Hadoop, is provided as an open source by Revolution Analytics, which was founded in 2007 as a spin-off from Yale University CS Department.

Pattern = " "

```
wc.map = function(., lines) { keyvalue( unlist( strsplit (x=lines,
    split=pattern)), 1)}  
wc.reduce = function (word, counts) { keyval (word, sum(counts))}  
Result = from.dfs ( mapreduce( input = "words.txt", output=NULL,  
    input.format = "text", map = wc.map, reduce = wc.reduce,  
    combine = T))
```



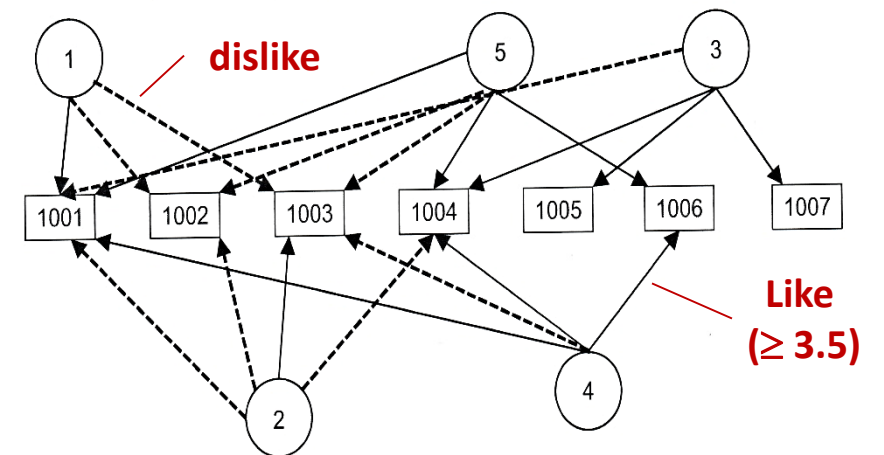
Big Data Analytics: Mahout

- 🌐 Mahout is an open source machine learning and predictive analytics intelligence library from Apache that use Hadoop.
- 🌐 Predictive analytics provide recommendation, classification and clustering.

```
Class MahoutRecommender {  
    public static void main (String [] args throws Exception {  
        DataModel model =  
            new FileDataModel (new File ("music_ratings.txt"));  
        UserSimilarity similarity =  
            new PearsonCorrelationSimilarity( model);  
        UserNeighborhood neighborhood =  
            new NearestNUserNeighborhood(3, similarity, model);  
        Recommender recommender =  
            new GenericUserBaseRecommender( model, neighborhood,  
            similarity);  
        List<RecommendedItem> recommendations =  
            recommender.recommend(1,2)    two recommendations for user 1  
        for (RecommendationItem recommendation : recommendations) {  
            System.out.println (recommendation);  
        }  
    }  
}
```

music_ratings.txt File in CSV format:
(userId, recordingId, userRating)

| | |
|--------------|--------------|
| 1, 1001, 5.0 | 4, 1001, 5.0 |
| 1, 1002, 3.0 | 4, 1003, 3.0 |
| 1, 1003, 2.5 | 4, 1004, 4.5 |
| 2, 1001, 2.0 | 4, 1006, 4.0 |
| 2, 1002, 2.5 | 5, 1001, 4.0 |
| 2, 1003, 5.0 | 5, 1002, 3.0 |
| 2, 1004, 2.0 | 5, 1003, 2.0 |
| 3, 1001, 2.5 | 5, 1004, 4.0 |
| 3, 1004, 4.0 | 5, 1005, 3.0 |
| 3, 1005, 4.5 | 5, 1006, 4.0 |
| 3, 1007, 5.0 | |



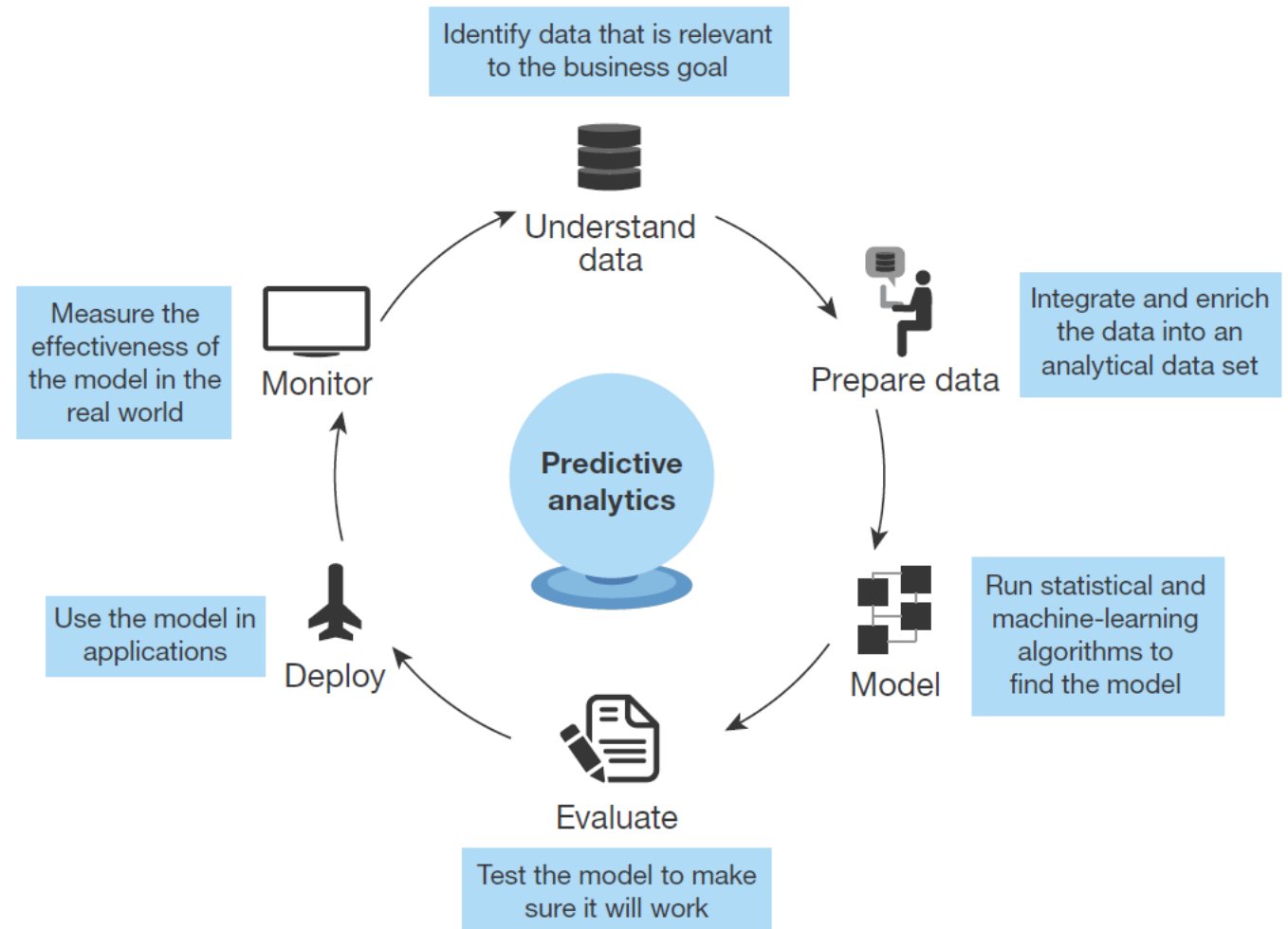
Big Data Analytics

Identify data from a variety of sources.

- Potentially valuable data often exists in multiple hard-to-access locations, both internally (data silos in enterprise applications) and externally (social media, government data, and other public or licensed data sources).



Wrangle the data.

- Data preparation for predictive analytics is a key challenge. Many users of predictive analytics spend more than three quarters of their time preparing the data: calculating aggregate fields, stripping extraneous characters, filling in missing data, or merging multiple data sources.





Big Data Analytics


Build a predictive model.

-  Predictive analytics include dozens of different statistical and machine-learning algorithms that data scientists can choose to run the best predictive model.
-  The best algorithm(s) to choose depend on the type and completeness of the data and the type of prediction desired.



Evaluate the model's effectiveness and accuracy.

-  Analysts run the analysis on a subset of the data called “training data” and set aside “test data” that they will use to evaluate the model
-  If the predictive model can predict the test data set, it is a candidate for deployment.

Use the model to deliver actionable prescriptions to your business peers.

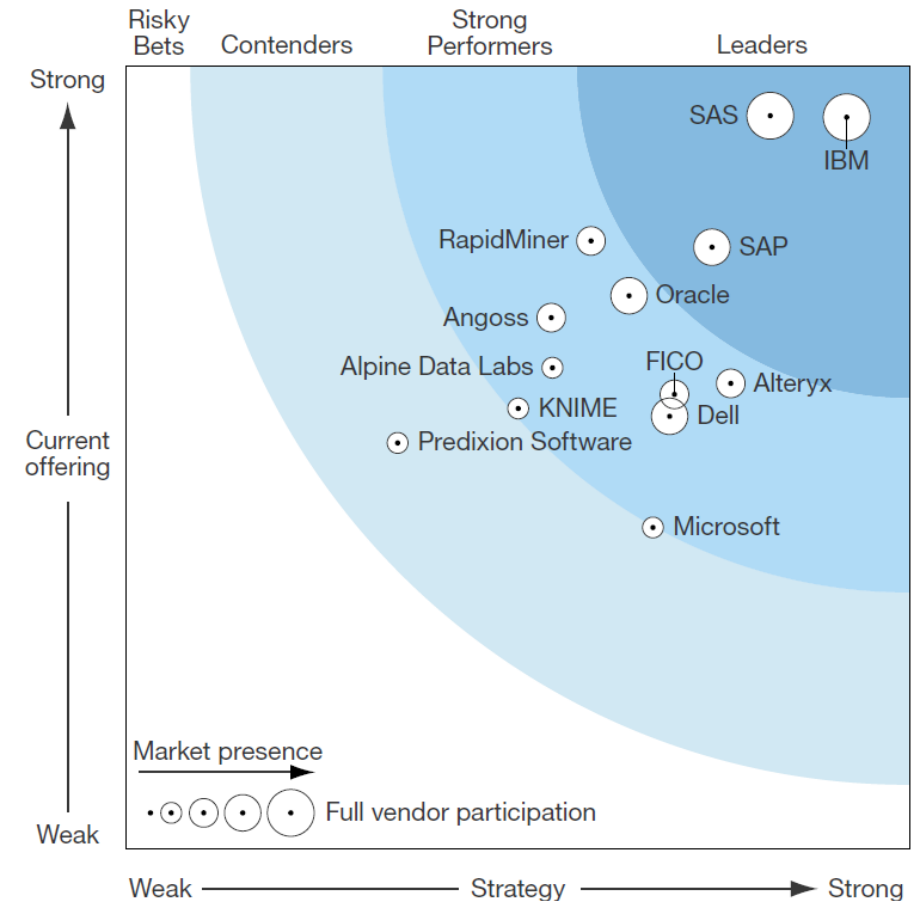
-  Business peers need to learn to trust in the predictions of models and those creating the models need to learn from their partners in the business what the most actionable insights may be.

Monitor and improve the effectiveness of the model.

-  Predictive models are only as accurate as the data fed into them, and over time they may degrade or increase their effectiveness.
-  If and when the model becomes less accurate, the model must be adjusted (e.g., by adjusting parameters in the algorithms) and/or seek additional data.

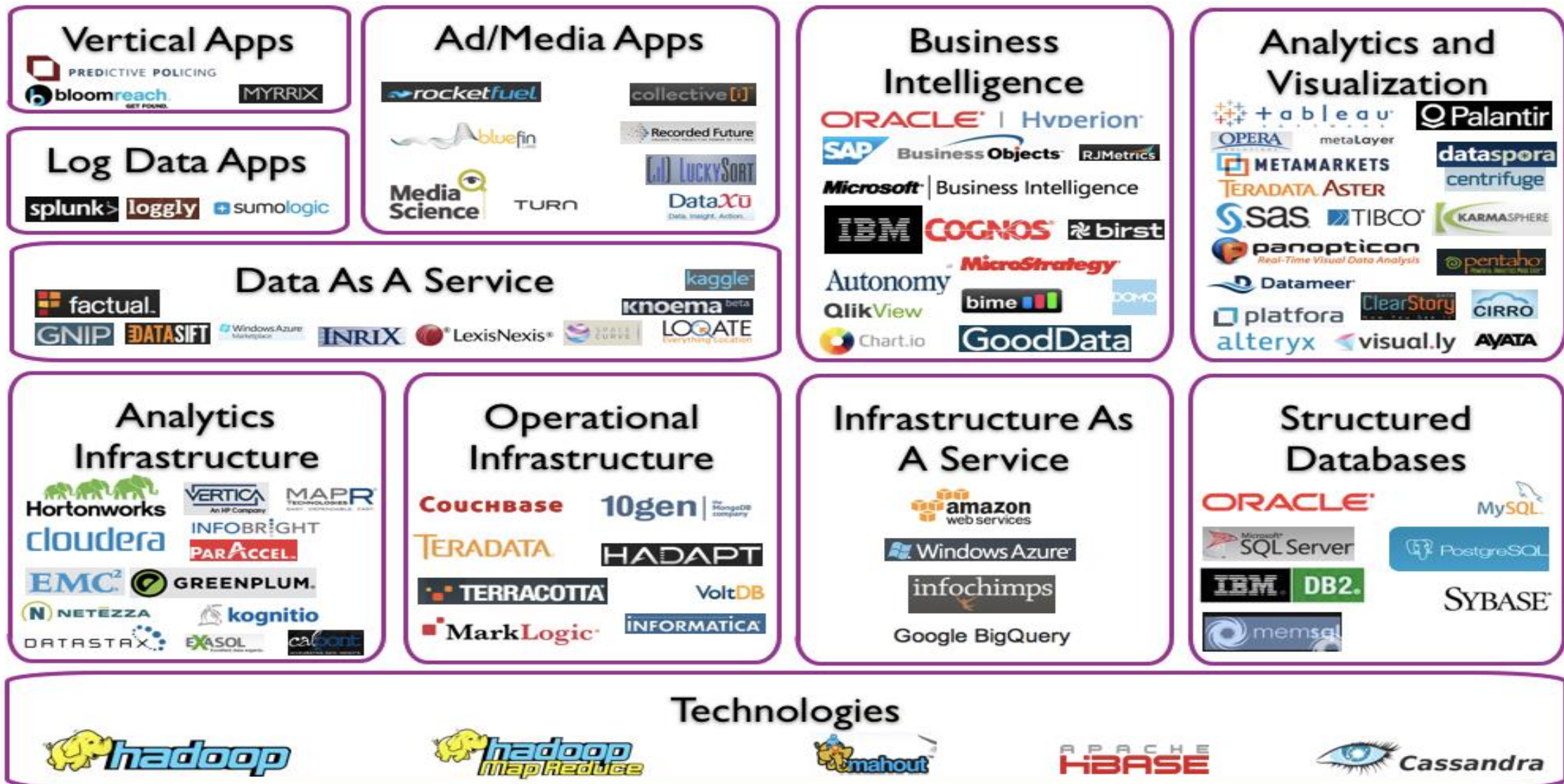
Big Data Analytics Trend

- Predictive analytics enable organizations to embed intelligence and insight into business processes.
 - Organizations in every industry are perking up to the value of predictive analytics.
 - Predictive analytics, however, has limited value unless the exposed insights can be deployed directly into software applications and business processes.
 - Information and predictive insights obtained from analytics must be linked to business decisions, process optimization, customer experience, or any other action
- Enterprises have lots of solid choices for big data predictive analytics solutions.
 - Predictive analytics has never been more relevant, and easier, than it is now.
 - Big data, gobs of compute power, and modern tools are making predictive models more efficient, accurate, and accessible to enterprises.
- Modern tools bring predictive power to more classes of users.
 - With a growth in demand, predictive analytics vendors are providing tools that lower the barrier to entry and increase appeal for those with less statistics skills.
 - Today's top predictive analytics tools can deploy their models or scoring engines into the applications where there is a need for insights. They enable application developers to use predictive analytics quickly and with increasing ubiquity in deployed applications.
 - API calls, web services, and predictive model markup language (PMMLs) are some of the methods that companies are using to seamlessly integrate predictions into their business.



The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015.

Big Data Landscape



Use Cases of Big Data Analytics



Web Log Analysis

- Click stream data show in details the kinds of things that people do on Web pages, such as what they looked at, but didn't buy.



Social Analytics

- Enterprises can find valuable information from the data in social networks and media.



Machine Generated Data Analysis

- Big data is often automatically generated by machines and sensors, from which important information can be mined.

Use Cases of Big Data Analytics



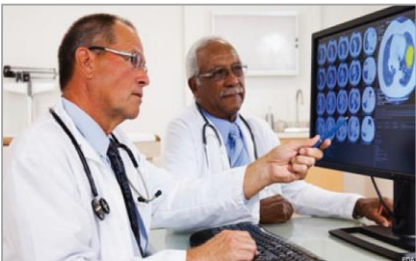
TXU Energy installed smart electric meters in customer homes and read the meter every 15 minutes. Based on an analysis of the metering data, it applies dynamic pricing to shape demand curve during peak hours. This eliminates the need for adding power generating capacity, saving millions of dollars for the company and saving customer expenditures as well.



T-Mobile USA has integrated data across multiple IT systems to combine customer transaction and interactions data in order to better predict customer defections. By leveraging social media data along with transaction data from CRM and billing systems, T-Mobile USA has been able to “cut customer defections in half in a single quarter”.



US Xpress collects about a thousand data elements ranging from fuel usage to tire condition to truck engine operations to GPS information, and uses this data for optimal fleet management and to drive productivity saving millions of dollars in operating costs.



Partners HealthCare, the largest healthcare provider in Massachusetts, is reusing ERM data—a mixture of structured and unstructured data—to dramatically speed up medical research by a factor of 10, and cut the cost by a factor of 5.

Case Study: PayPal

🌐 Company

- 🔍 Global e-commerce business allowing payments and money transfers made through Internet.

🌐 Role of Global Business Analytics Team

- 🔍 Managing Down: Ensure connection between the analysis they do and the actions the company takes. Work closely with business people for right questions and right interpretation of findings.
- 🔍 Managing Up: Establish themselves as thought partners, not data providers, to the executive, and translate analytical insights into actionable recommendations.

🌐 Analytics Team Members

- 🔍 Business analysts with a mix of technical and business skills. Most having MBAs in addition to data analysis skills.

🌐 Project Examples

- 🔍 Analysis of customer behaviors and interactions for improving products and marketing, analysis of the impact of website redesign, analysis of the effect of promotional pricing, diagnosis of revenue leakages, analysis of the impact of risk management policies on customers, etc.



Veronika
Belokhvostova,
Head of Global
Business Analytics
at PayPal

Case Study: Sears

Company

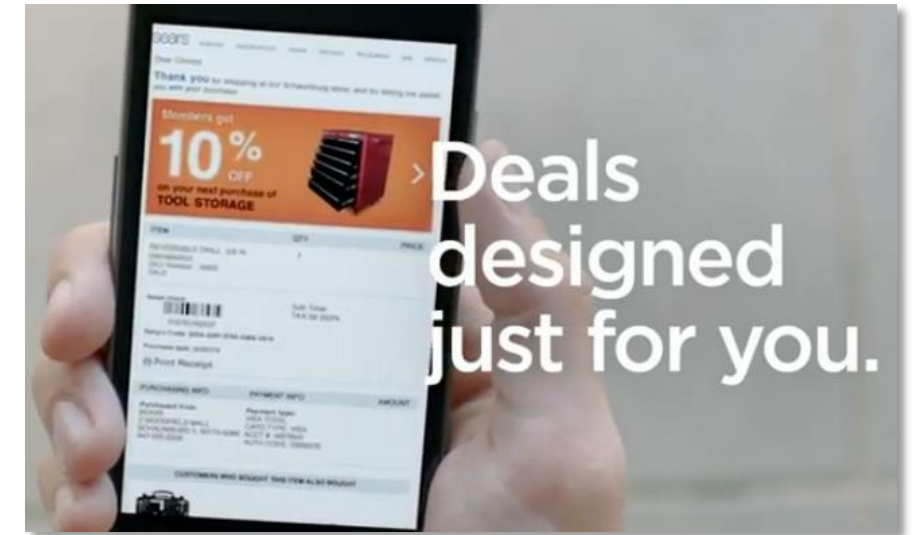
- American chain of department stores

Challenge

- Decided to generate greater value from the huge amounts of customer, product and promotion data collected from its stores.
- Took 8 weeks, due to highly fragmented databases and data warehouses, to generate personalized promotions, at which point many of them were no longer optimal.

Solution

- Set up a Hadoop cluster in 2010, and used it to store incoming data from its stores and to hold data from existing data warehouses.
- Conducted analyses directly on the cluster, with the processing time reduced from 8 to 1 week, and still dropping.
- Got help from Cloudera initially, but over time internal IT and analysts became comfortable with the new tools and methods.



Big Data Platforms on Cloud



- Amazon offers flexible storage models that provide instant dynamic scaling; Elastic Map Reduce, which is a hosted, scalable Hadoop service; DynamoDB, a NoSQL database; and tools to help with analytics.



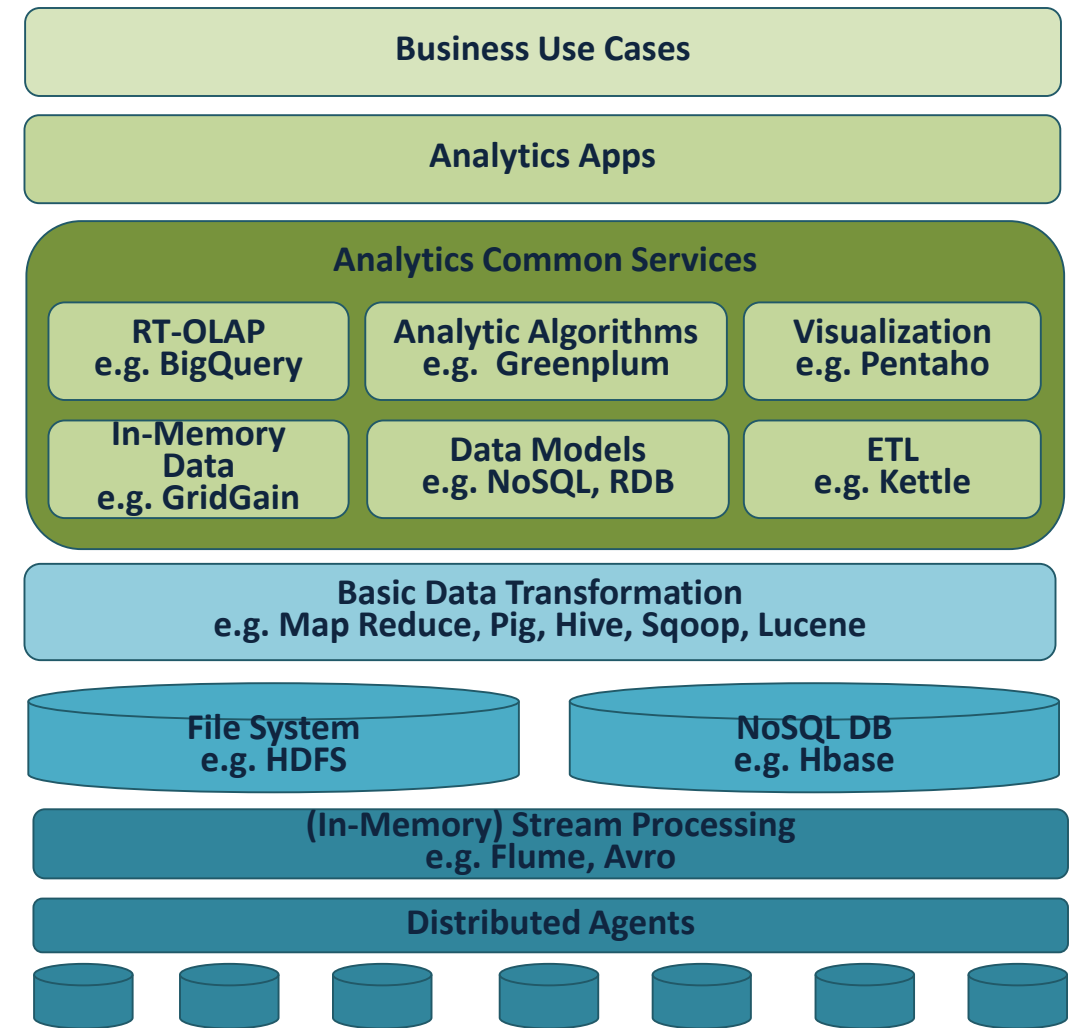
- BigQuery provides big data analytics services and Prediction API provides machine learning services.
- Its cloud application hosting service, AppEngine, also offers a MapReduce facility.
- Google invested in DNAnexus, a company specializing in storage and analysis services for DNA sequencing.



- Hortonworks is helping Microsoft develop its own Hadoop-based offering on Windows Server and Windows Azure.
- Redmond plans to contribute its adaptations back to the Apache Hadoop project, which means anybody will be able to run a purely open source Hadoop on Windows.

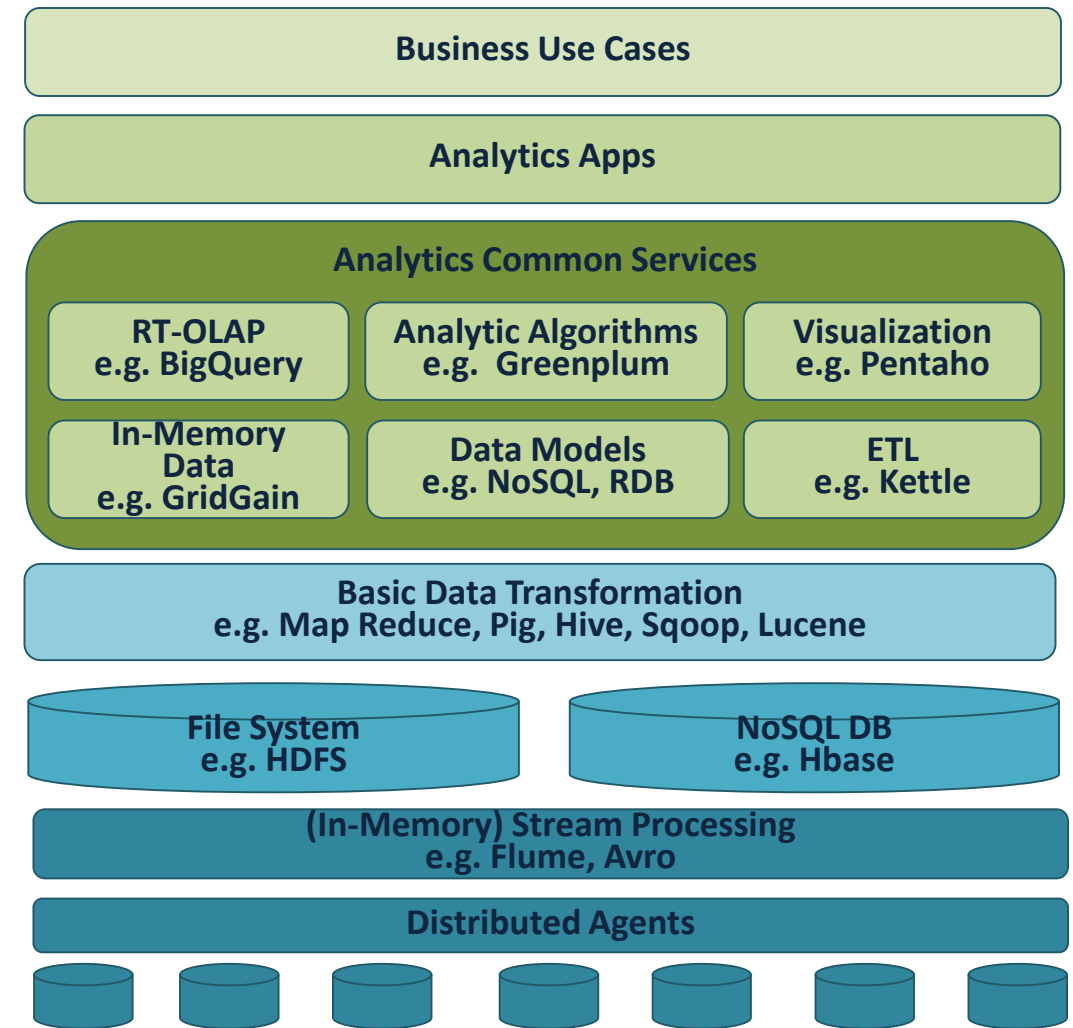
SaaS based on Big Data Analytics

- Use low-cost, open-source tools in pilots to demonstrate the feasibility of big data projects.
 - R, the open source programming language for statistics and predictive analytics
 - API libraries available to prepare data and build predictive models using Java, Python, and Scala. (Apache Mahout and WEKA have Java APIs. Apache Spark Mllib includes APIs for Java, Python, and Scala.2 Python developers can use NumPy and SciPy to prepare data and build predictive models.)
- Explore the increasing number of public datasets now available through open APIs.
- Produce a resource plan that identifies big data skill gaps. Look for business-savvy analysts (especially data scientists) and analytics-savvy business leaders who can work together to find what business should do based on analytic results and then do it.
- Assess resource needs for information infrastructure and identify technical gaps when supporting big data solutions.



SaaS based on Big Data Analytics

- For enterprise SaaS, big data analytics requires a data-savvy business models that leads to competitive advantage.
- Keep the business process transparent; it is key to successful big data applications.
- Educate process owners about potential big data opportunities now readily available through start-small, cost-effective analytics tools and techniques.
- The value delivered from an investment in big data analytics must be visible and measureable.



Data Scientists

How to Find the Data Scientists You Need

1 Focus recruiting at the “usual suspect” universities (Stanford, MIT, Berkeley, Harvard, Carnegie Mellon) and also at a few others with proven strengths: North Carolina State, UC Santa Cruz, the University of Maryland, the University of Washington, and UT Austin.

2 Scan the membership rolls of user groups devoted to data science tools. The R User Groups (for an open-source statistical tool favored by data scientists) and Python Interest Groups (for PIGgies) are good places to start.

3 Search for data scientists on LinkedIn—they’re almost all on there, and you can see if they have the skills you want.

4 Hang out with data scientists at the Strata, Structure:Data, and Hadoop World conferences and similar gatherings (there is almost one a week now) or at informal data scientist “meet-ups” in the Bay Area; Boston; New York; Washington, DC; London; Singapore; and Sydney.

5 Make friends with a local venture capitalist, who is likely to have gotten a variety of big data proposals over the past year.

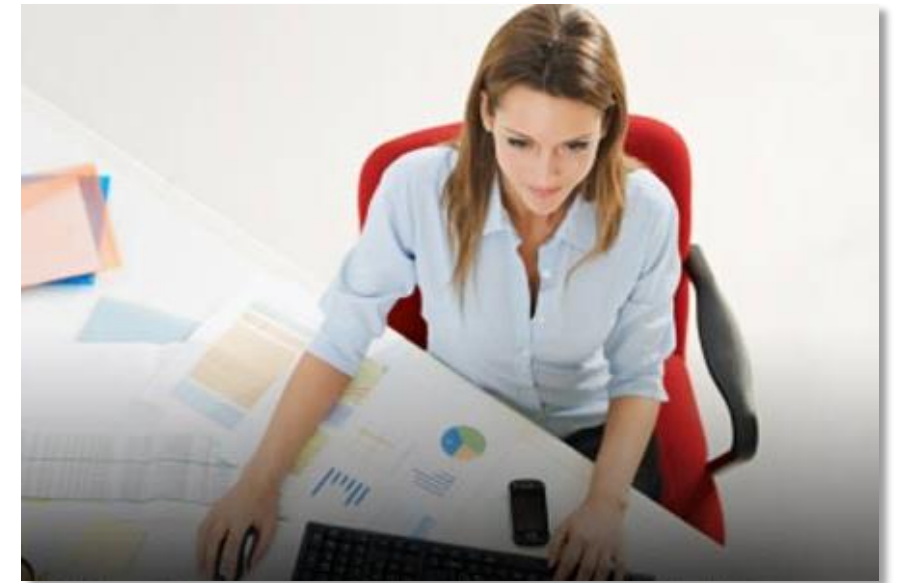
6 Host a competition on Kaggle or TopCoder, the analytics and coding competition sites. Follow up with the most-creative entrants.

7 Don’t bother with any candidate who can’t code. Coding skills don’t have to be at a world-class level but should be good enough to get by. Look for evidence, too, that candidates learn rapidly about new technologies and methods.

8 Make sure a candidate can find a story in a data set and provide a coherent narrative about a key data insight. Test whether he or she can communicate with numbers, visually and verbally.

9 Be wary of candidates who are too detached from the business world. When you ask how their work might apply to your management challenges, are they stuck for answers?

10 Ask candidates about their favorite analysis or insight and how they are keeping their skills sharp. Have they gotten a certificate in the advanced track of Stanford’s online Machine Learning course, contributed to open-source projects, or built an online repository of code to share (for example, on GitHub)?



Thomas Davenport and D. Patil, Data scientist: the sexiest job of the 21st century, Harvard Business Review Oct. 2012.