Big Data Overview

Dec.08,2016 FANG, Jiachao

Content

- Story of Big Data
- Process of Big Data Projects
- Big Data Infrastructure
- Case Study: Ford

Story of Big Data

Embryonic Stage

1980 - 1996

Key word: Volume
In this stage, big data only means **massive data**, do not involve any processing technology or storage methods.

- 1980, the term "**Big Data**" first used by futurist, Alvin Toffler
- The development of communication industry leads to increase of **information flow**

Developing Stage

1996 - 2006

Key word: 3Vs
In this stage, **3Vs(Volume, Velocity, Variety)** became the generally-accepted three defining dimensions of big data.

- Maturity of **data mining**[1] theories and techniques
- 2003, the break out of **unstructured data**[2] drives the development of data processing

Flourishing Stage

2006 - now

Key word: Big Data Era
In this stage, with the maturity of all kinds of big data techniques, big data is widely applied in almost all fields.

- The popularity of smartphones leads to rapid growth of **mobile data**
- Big data application has been grown from theory research to **applied period**

1 Data Mining is a computational process of discovering patterns in large data sets
2 Unstructured data refers to information that either does not have a pre-defined data model or is not organized in a predefined manner

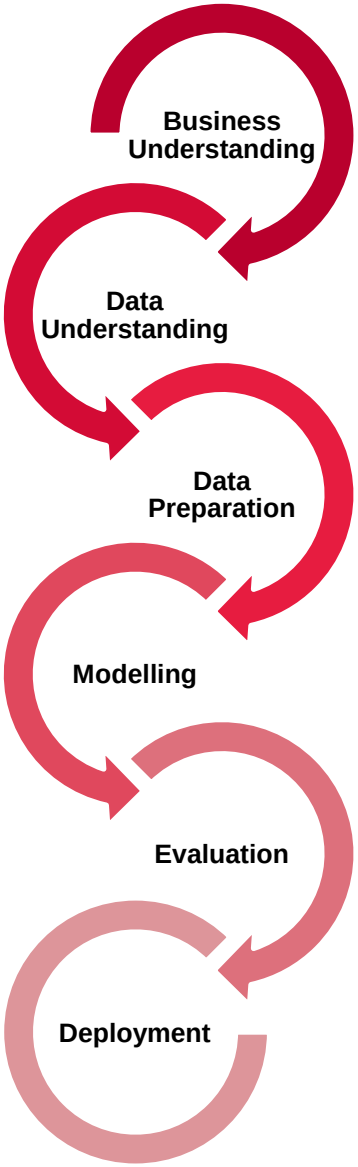
Process of Big Data Projects in Industry

1. Business Understanding

- Understand business objectives
- Assess the current situation
- Create data mining goals
- Establish data mining plans

2. Data Understanding

- Initial data collection
- Data load and integration
- Examine data properties
- Examine data quality



3.Data Preparation

- Data selection
- Data cleaning
- Data construction
- Data formatting

4. Modeling

- Select suitable modelling techniques
- Use test scenario to validate the model
- Model creation
- Model assessment

5. Evaluation

- Evaluate model results in the context of business objectives
- New business requirements may be raised
- “Go or no-go” decision must be made

6. Deployment

- Create a report or a repeatable data mining process
- Create deployment ,monitoring and maintenance plan
- Create a summary project report

Basic Steps of Big Data

Determine Data Sources → **Data Collection** → **Data Storage& Processing** → **Data Analytics** → **Data Visualization**

Determine Data Sources

Social Media Networks

Sensors in Cars

Data Collection

Sentiment API

*Sentiment analysis

ETL Tools

*Extract-Transform-Load

Data Storage& Processing

Framework

Data Processing Tools

Data Analytics

Statistical Tools

Data warehouse

Data Visualization

Applications & Visualization

Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice machine, bluedata, jethro

Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, Qubole, xplenty

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, memsql, nuoDB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

Graph Databases
neo4j, GIRAPH, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, kognitio, dremio

Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

Data Transformation
alteryx, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData

Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk, DATADOG, rocana, Anodot

Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, Qumulo

App Dev
apigee, CASK, Kaseya IO, Typesafe, CONCURRENT

Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, inter|ana

Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, MODE, plotly, ADATAO, dataiku, tonian, DOMINO, sense, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Roambi, GOMDATA, Qlik, CHARTIO

BI Platforms
Power BI, amazon web services, DOMO, Wave Analytics, GoodData, birst, platfora, looker, atscale, ARCADIA, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk>, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
NETBASE, DATASIFT, tracx, bitly, synthesio, bottlenose, simplereach

Real-Time
amazon web services, METAMARKETS, confluent, DATATORRENT, dataArtisans

Machine Learning
Azure Machine Learning, amazon web services, H2O.ai, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfl.ash

Speech & NLP
NarrativeScience, api.ai, NUANCE, Gridspace, semantic machines, cortical.io, MindMeld, IDIBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient, VIV, vicarious, nervana, Numenta, nara, SCALE INFERENCE, HyperScience, MetaMind, clarifai, NEXTO, Geometric Intelligence

Search
hp Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

Data Services
Mu Sigma, OPERA, DATASCIENCE, DATA SCIENCE, kaggle, datascope, DataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

SMB / Commerce
Google Analytics, AMPLITUDE, RJMetrics, BLUECORE, sumall, granify, Airtable, retention science, custora

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blueyonder, kahuna, Lattice, SAILTHRU, persado, infer, 6sense, AVISO, ACTIONIQ, QUANTIFIND, ENGAGIO

Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, STELLAService, NGDATA, Preact, Wise.io, DigitalGenius, appuri, fuse|machines

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

Ad Optimization
MediaMath, Integral Ad Science, OpenX, rocketfuel, theTradeDesk, Adgorithms, LiveIntent, dstillery, DataXu, Appier, TAPAD

Security
CYLANCE, CounterTack, cybereason, AREA 1 SECURITY, ThreatMetrix, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

Vertical AI Applications
X.ai, facebook, Clara, KASISTO, lumiata

Publisher Tools
outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo

Govt/ Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, nesfinance, LendUp, Kabbage, tidemark, INSIKT, zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient

Education/ Learning
KNEWTON, Clever, declara, PANORAMA, knowre

Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, sas, hp Autonomy, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Apache Kafka, Flink, CDAP

Query / Data Flow
SLAMDATA, HIVE, APACHE DRILL, Google Cloud Dataflow

Data Access
cassandra, APACHE HBASE, accumulo, mongoDB, CouchDB, riak, SciDB, kafka, nifi, OPENTSDB

Coordination
talend, ZOOKEEPER, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mllib, mahout, Aerosolve, Apache SINGA, MADlib, Caffe, learn, CNTK, TensorFlow, FeatureFu, VELES, WEKA, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
JAWBONE, GARMIN, practice fusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara, AUGURY, estimote

Financial & Economic Data
Bloomberg, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, estimize, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location/People/Entities
GARMIN, foursquare, InsideView, esri, STREETLINE, CARTODB, factual, Place IQ, Crimson Hexagon, placemeter, BASIS, Sense 360

Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools

GA, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

FIRSTMARK

Big Data Framework

What exactly is a big data framework?
It is a combination of technologies and methodologies that is used to **transform big data** in its **raw form** into **refined data** governed by a mature framework that can continuously be used for virtually any big data application — whether it is a batch report, a near-real time data stream analysis or a dashboard.

	1 hadoop	2 Spark	3 APACHE STORM™ Distributed · Resilient · Real-time
Differences			
Data Processing Models	› best suited for batch processing	› batch processing and real time processing	› supports micro-batching
Development	› written in Java and implemented using Apache pig	› Implemented by Scala tuples, a bit difficult to implement over java	› uses DAG's on every node and data transfer through Storm tuples
Scenarios	› Batch data processing › distributed computing model based on RPC	› Multiple operations of specific data sets	› Offline analysis of massive data › Large scale search of web info

Similaritie

Hadoop, Spark and Storm are preferred choice of frameworks amongst developers for big data applications (based on the requirements) because of their **simple implementation methodology**.

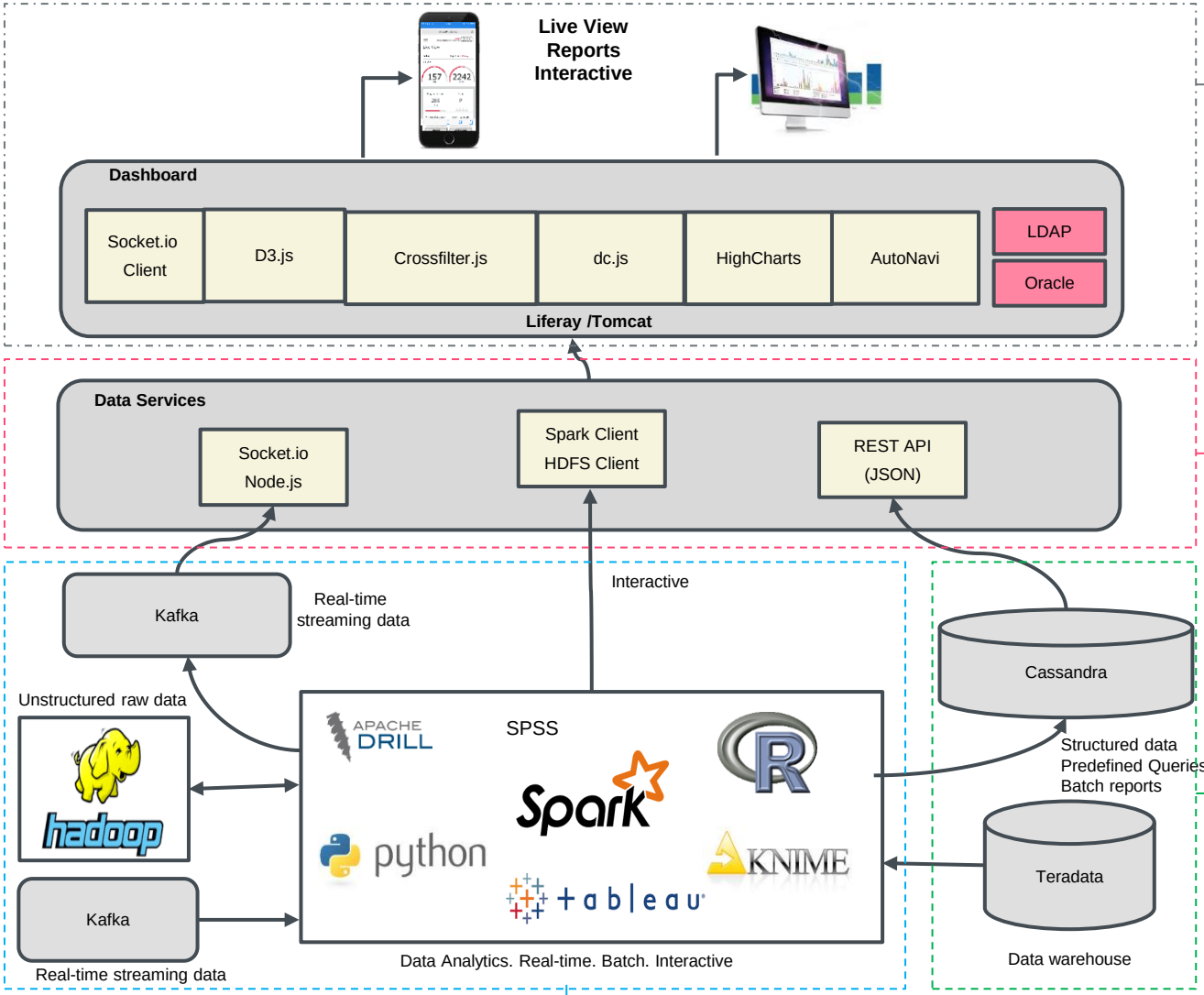
- open source processing frameworks
- real time BI and big data analytics
- Implemented in **JVM** based programming languages
- Provide fault tolerance and scalability

Stat Tools for Big Data

- **Statistical tools are used for data processing, statistics and data visualization**
- **The right choice of statistical tools will largely increase the efficiency of data analysis**
- **SAS, R and Python are most frequent used among all**

	1	2	3
	sas e-Intelligence	R	python
Advantages	› Market leader in **commercial** analytics space › Huge array of statistical **functions** › Quickly learnt	› Open source counterpart of SAS › Traditionally been used in **academics and research** › Latest techniques get **released quickly** › A very **cost-effective** option.	› An open source **scripting language** › Apply to almost any statistical operation / model › **Strong performance** in operations on structured data
Disadvantages	› **Expensive** › not always enriched with latest statistical functions › Low graphical capacity	› **Difficult to learn** ***** Documentation is sometimes patchy and terse, and impenetrable to the non-statistician	› **Smaller pool of Python developers** compared to other languages, such as Java › **Lack** of true **multiprocessor** support

Big Data Infrastructure & Team



Data Product Manager

- Analysis and scoop out customer needs
- Design applications to provide users with data visualization

Data Scientist

- Algorithm optimization

Data Modeler

- Data modeling
- Improve data utilization

Business Analyst

- Propose solutions according to business requirement

Data Warehouse Architect

- Data warehouse construction

Data Developer

- Data cleaning, data formatting
- Respond data analyst's requirement

Data Mining Engineer

- Excavate valuable information from massive data

Data Developer

- System building, maintenance & update;
- Keep platform stable

Data Architect

- Design underlying platform architecture
- Plan technology roadmap

Operations

- Daily operations of platform

Case study: Ford

How Big Data Brought Ford Back from the Brink

Background

2006

- Ford's in trouble
- Report of a **$12.6 billion** loss

2007

- Applications of **Data Science**

2009

- Posted profits for the first time in 4 years
- Launched 25 new vehicle lines
- Sold **2.3 million** cars

2013

- 2013 INFORMS Prize for Company-Wide Efforts **in Analytics & Data Science**

Case study: Ford

Improving Supply & Demand

Data Sources

What is being built

Public Sentiment

Sales Data

Dealership location Info

…

Smart Inventory Management System

Data Integration

Techniques

Text mining/ data mining

what consumers want that is not available

Predictive analytics*

make decisions on new car features

* Using analytical skills to analyze current and historical facts to make predictions about future or otherwise unknown events.

Results

Dealer Stock/ Car Specification

Supply Chain

Consumer Demands

Case study: Ford
Improving Efficiency on the Factory Floor

Before

high cost of producing prototypes

Solution

Prototype Optimization Model

- Computing the **minimum number of vehicles** required to perform the **maximum amount of tests**

After

- An estimated **$12 million** in savings first used on the European Transit vehicle
- An estimated **$250 million** in savings if rolled out the whole company

Case study: Ford

Anticipating Future Energy Supply and Demand

Data Sources

- **performance data** from sensors in electric cars

Where customers plug in

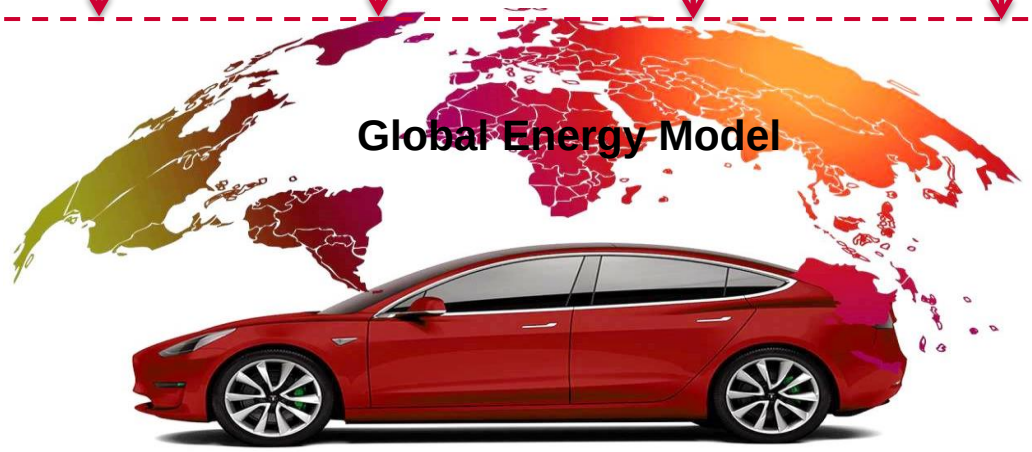
How many gas miles they drove

How many electric mile

How often they take trips

...

Solution



Results

- Detailed **reports for drivers** about car performance and driving conditions

- Developing vehicles using a range of **new fuel technologies**

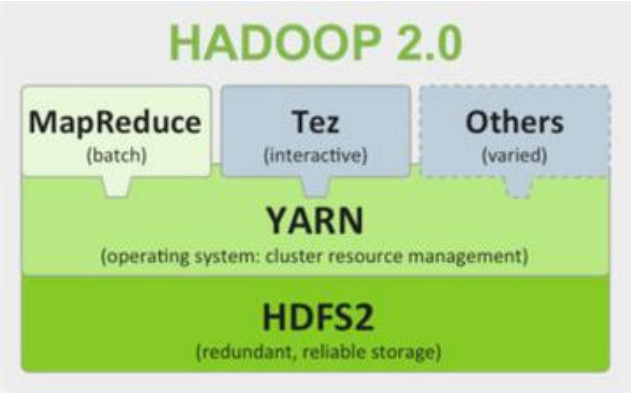
- Combined with **internal process** and knowing where to improve

Case study: Ford

Ford's Big Data Strategy

Data Sources

Analytics Platform



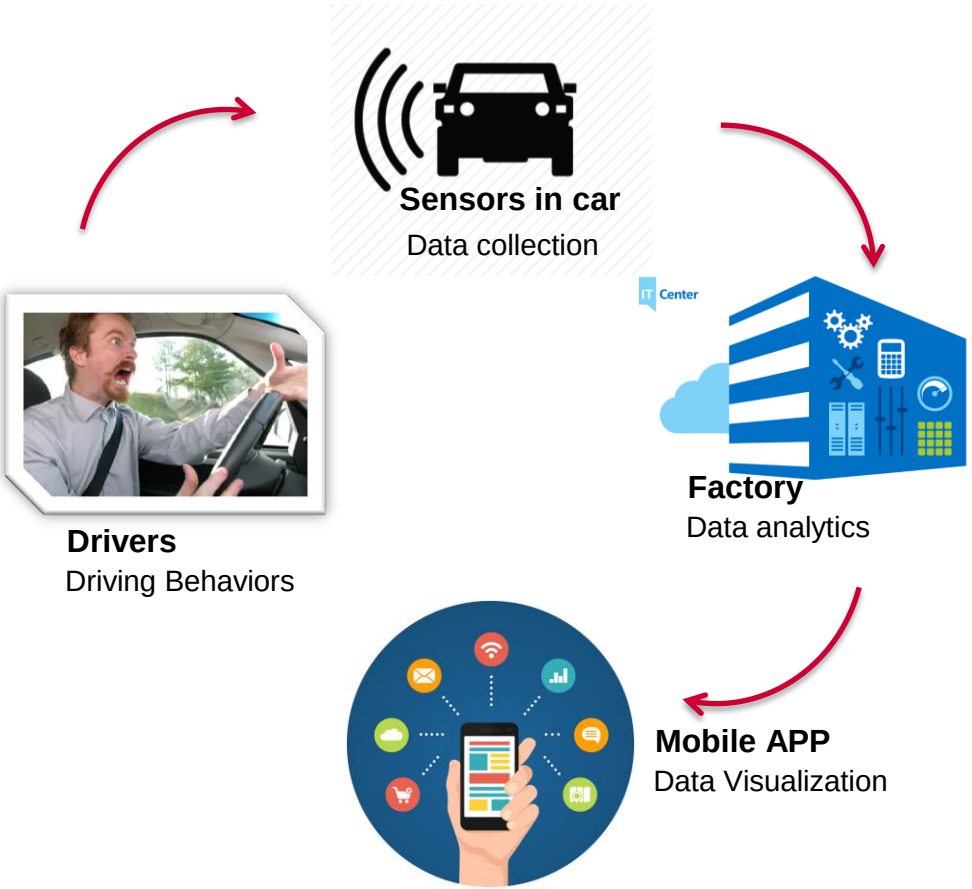
Tools

- R :statistical analysis

Methods

- sentiment analysis; real-time analysis; data mining; text mining

Instance: Behavior Monitoring



- Thank you!