# Introduction to Big Data :

Name: อวยชัย    ภิรมย์รื่น

Tel.   : 086-813-5354

e-mail : p.Auoychai@gmail.com

# Big Data
## What is it?
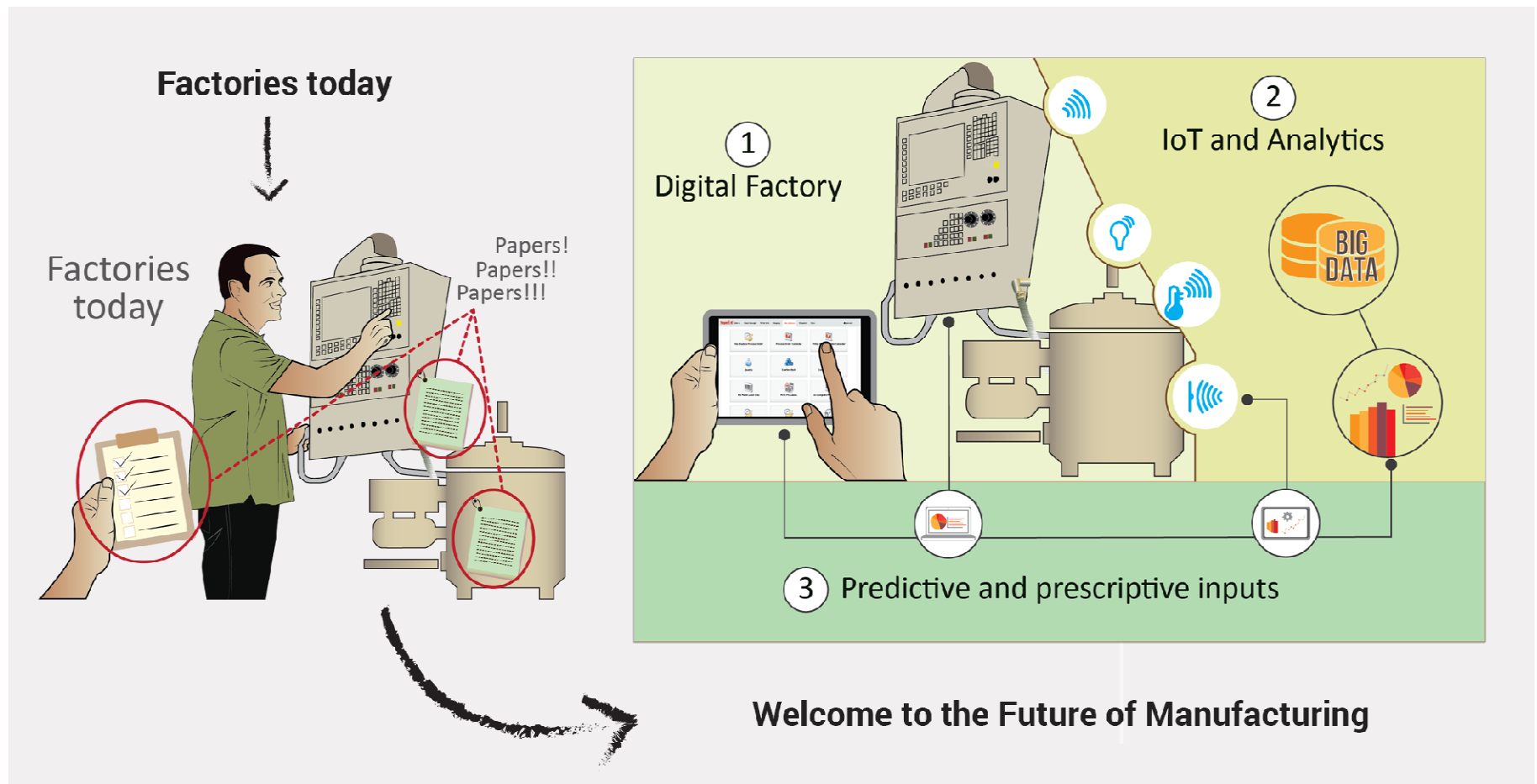
# Question : BIG DATA

*@* อะไรคือ BIG DATA

   *@* ทำไมเราต้องสนใจ BIG DATA
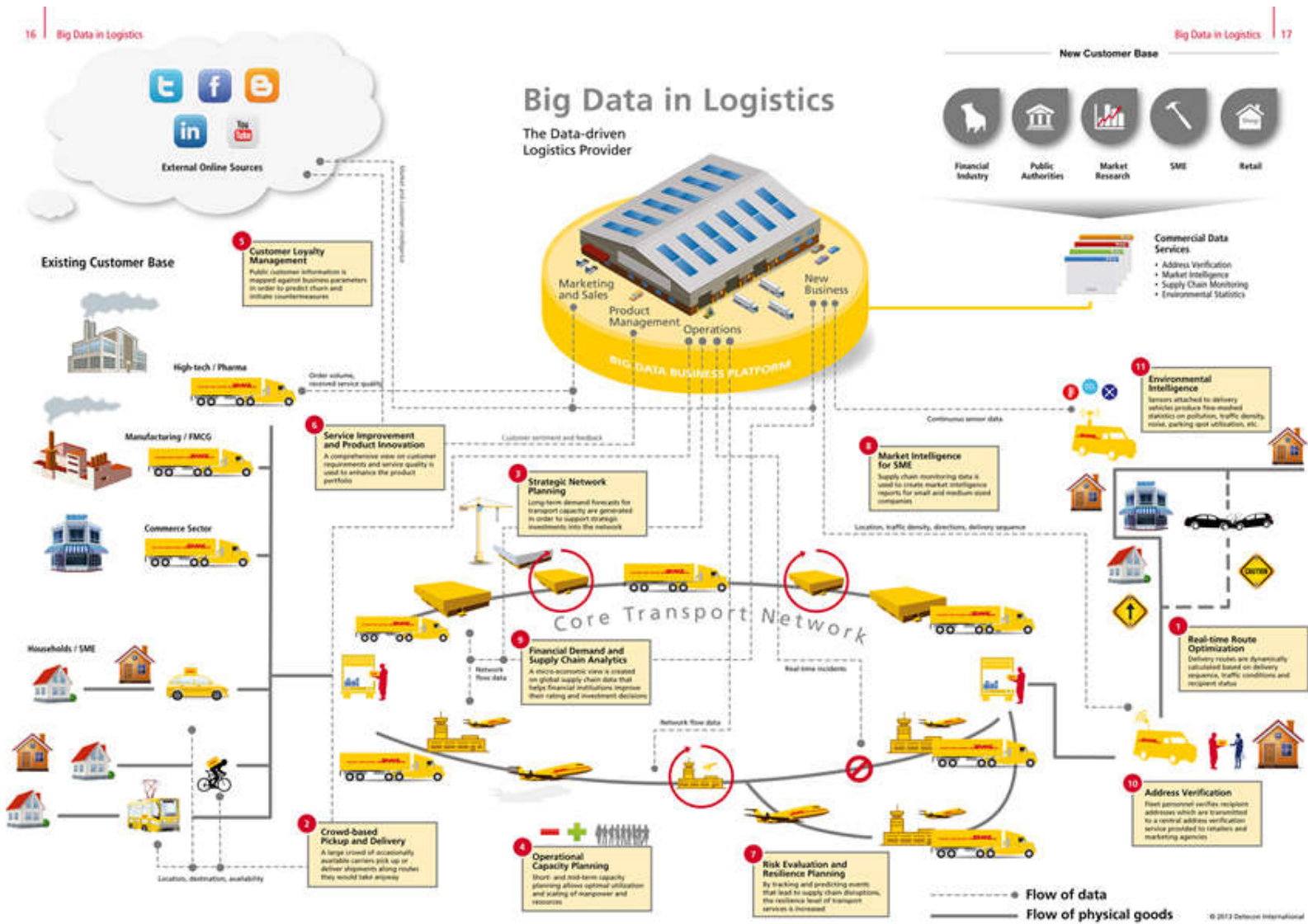
*@* เราต้องการอะไรจาก BIG DATA

# # Why need BIG DATA : Industry 4.0

# Why need BIG DATA : Modern Coffee Shop

# # Why need BIG DATA : Logistic



https://i0.wp.com/cerasis.com/wp-content/uploads/2016/11/2016-11-14.png

# Why need BIG DATA : Traffic

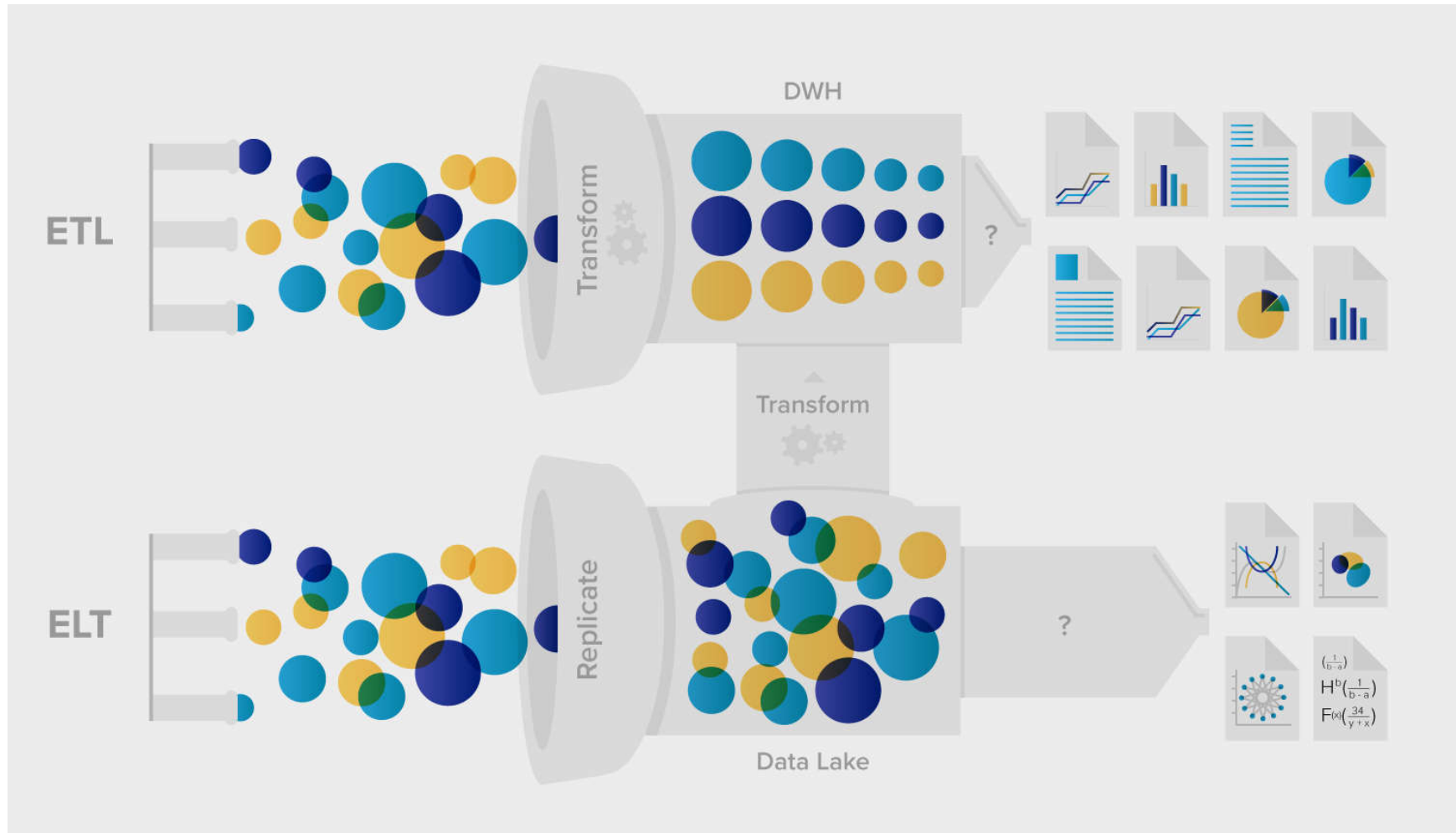# Why need BIG DATA : wholesale

# Why need BIG DATA : hospital



https://www.meed.com/pictures/1240x826/7/5/0/2003750_big-data-healthcaredreamstime_m_49010041.jpg

# Why need BIG DATA : Banking



https://cdn.computerworlduk.com/cmsdata/slideshow/3657529/Bribe_money_cash_bribe_thumb800.jpg

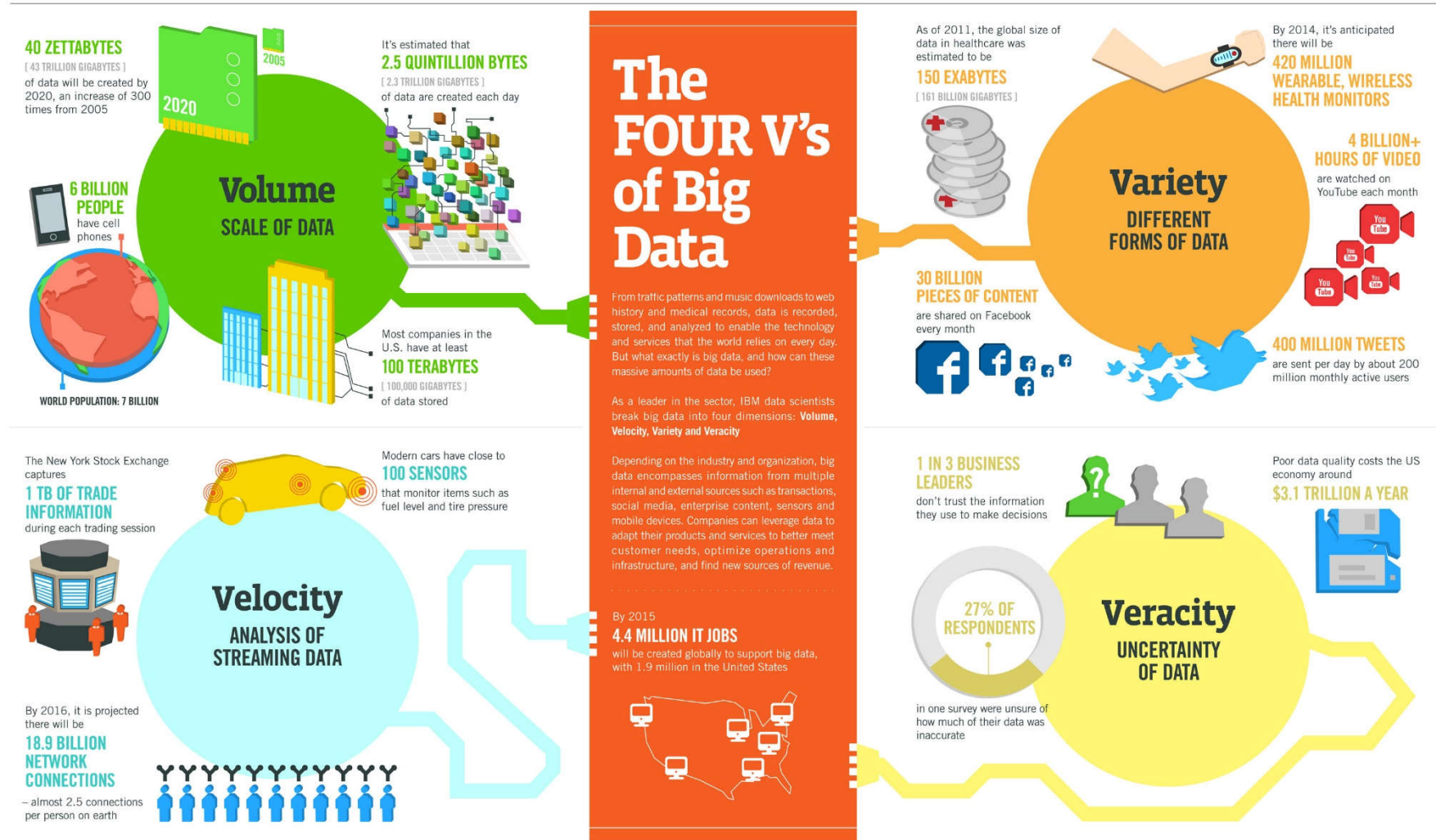# Why need BIG DATA : Data Warehouse



https://cdn.computerworlduk.com/cmsdata/slideshow/3657529/Bribe_money_cash_bribe_thumb800.jpg

# BIG DATA มาจากไหน

Activity Data
Conversation Data
Photo and Video Image Data
Sensor Data
The Internet of Things Data

# # BIG DATA Definition :



**The FOUR V's of Big Data**

**40 ZETTABYTES** [ 43 TRILLION GIGABYTES ] of data will be created by 2020, an increase of 300 times from 2005

2005 / 2020

**6 BILLION PEOPLE** have cell phones

WORLD POPULATION: 7 BILLION

**Volume** SCALE OF DATA

It's estimated that **2.5 QUINTILLION BYTES** [ 2.3 TRILLION GIGABYTES ] of data are created each day

Most companies in the U.S. have at least **100 TERABYTES** [ 100,000 GIGABYTES ] of data stored

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [ 161 BILLION GIGABYTES ]

**30 BILLION PIECES OF CONTENT** are shared on Facebook every month

**Variety** DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

**4 BILLION+ HOURS OF VIDEO** are watched on YouTube each month

**400 MILLION TWEETS** are sent per day by about 200 million monthly active users

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

**Velocity** ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

**1 IN 3 BUSINESS LEADERS** don't trust the information they use to make decisions

**27% OF RESPONDENTS** in one survey were unsure of how much of their data was inaccurate

**Veracity** UNCERTAINTY OF DATA

Poor data quality costs the US economy around **$3.1 TRILLION A YEAR**

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

https://www-01.ibm.com/software/data/bigdata/

# The Sizing of BIG DATA



http://images.rigzone.com/images/home/article/infog_140631.jpg

# The challenge of BIG DATA :
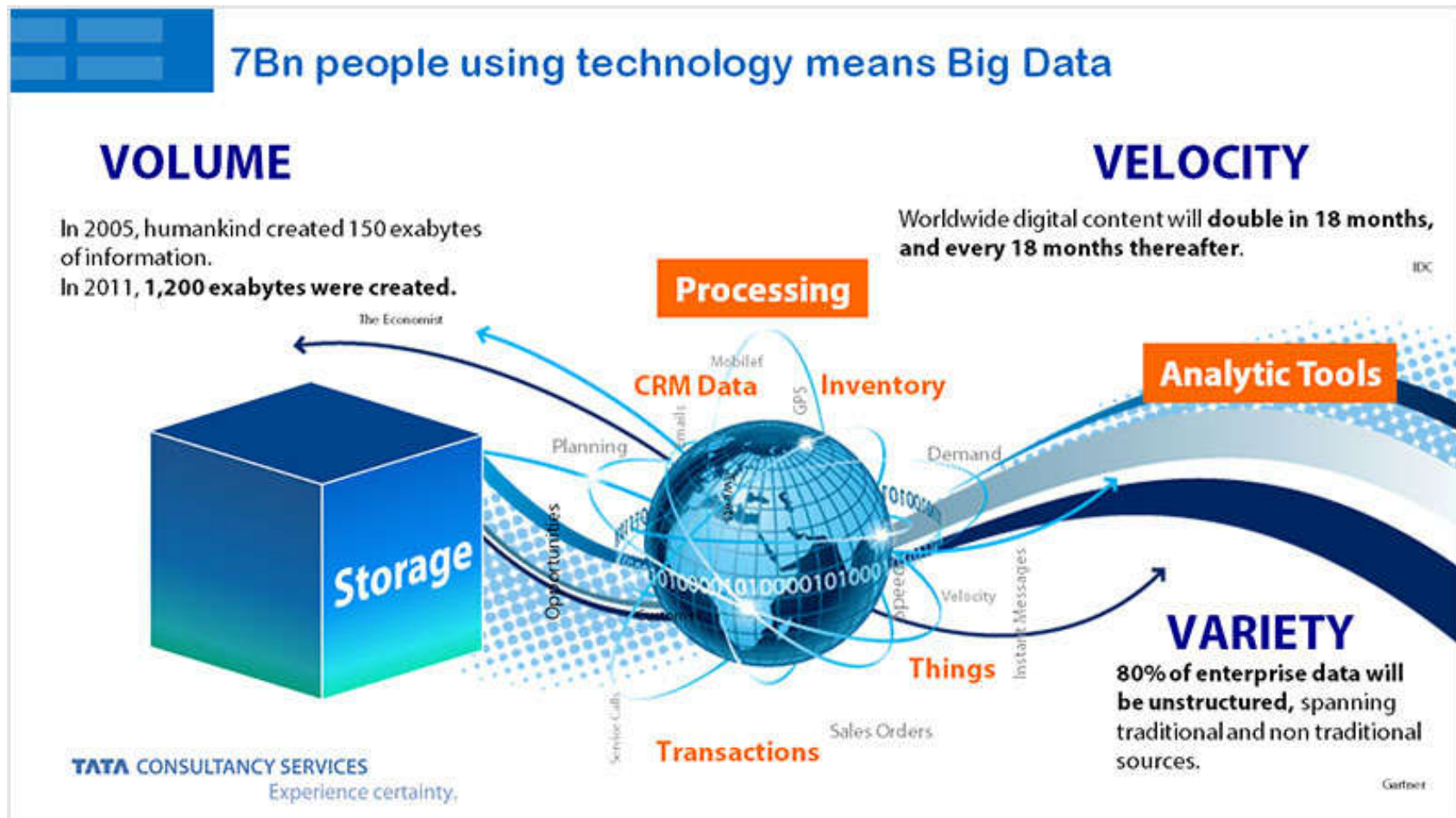
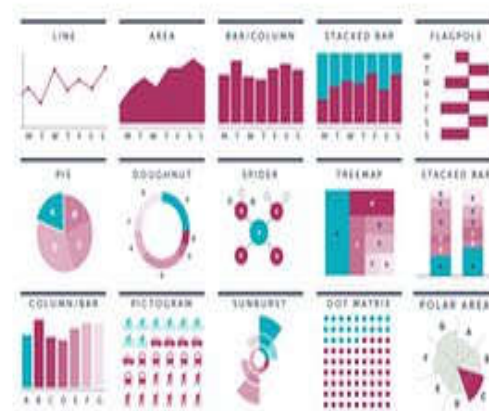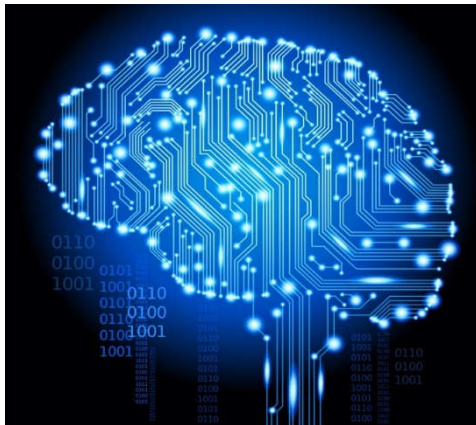# The things cover to BIG DATA Technology :
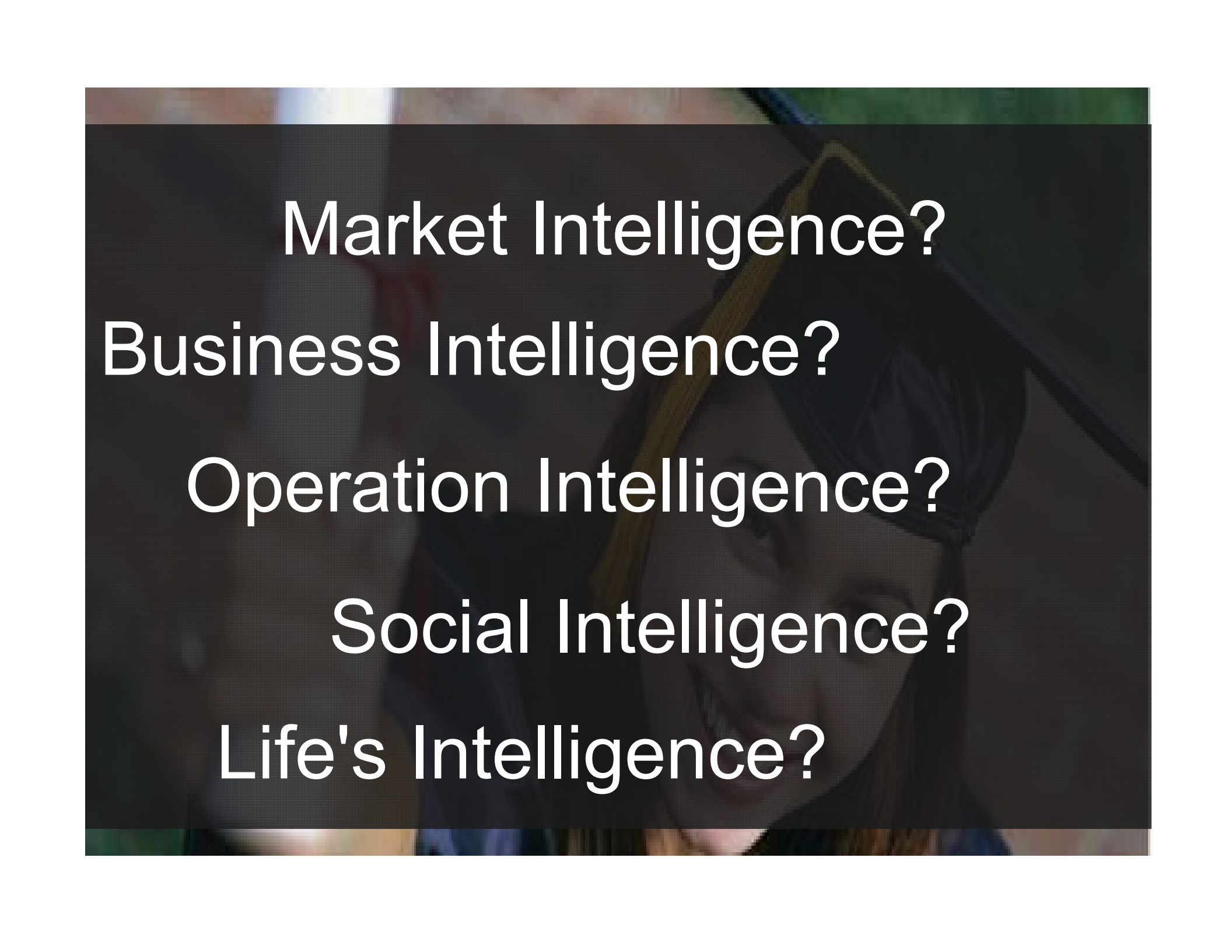


**Big Data Technology**

# Big Data = Big + Data - Big

**Big Data is not about the *size* of the data,
it's about the *value* within the data.**

So, what is
*value?*

Operation innovation

Operation efficiency

User satisfaction

Benefits of insights

Business innovation

Business efficiency

Subscriber personalization

select – combine – analyze

Business data

Session data

Signaling data

Management data
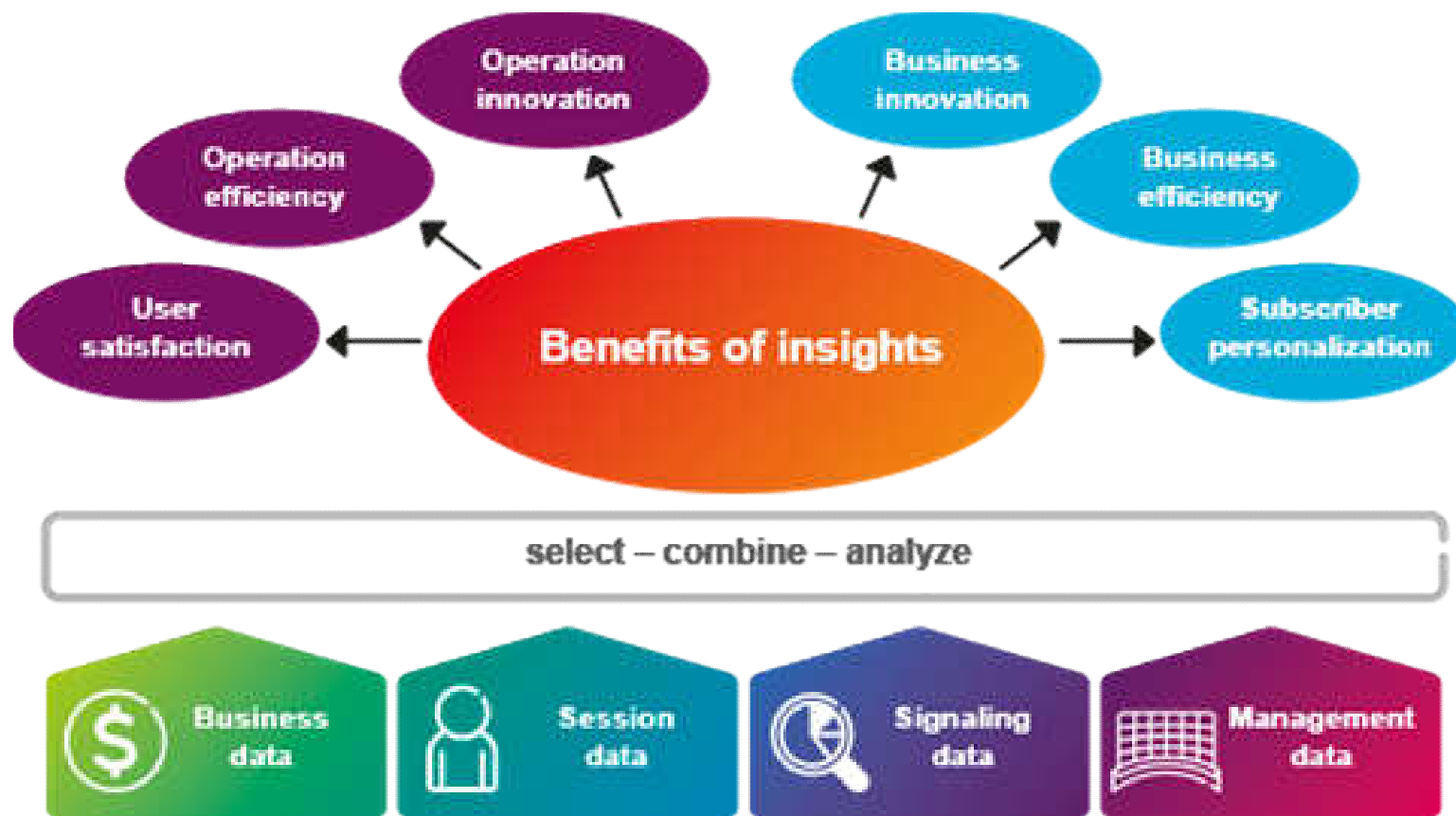
Market Intelligence?

Business Intelligence?

Operation Intelligence?

Social Intelligence?

Life's Intelligence?

# Big Data Landscape

## Vertical Apps
PREDICTIVE POLICING
bloomreach GET FOUND.
MYRRIX

## Log Data Apps
splunk> loggly sumologic

## Data As A Service
factual. kaggle knoema beta
GNIP DATASIFT Windows Azure Marketplace INRIX LexisNexis SPACE CURVE LOQATE Everything Location

## Ad/Media Apps
rocketfuel collective[i]
bluefin Recorded Future
Media Science TURN LuckySort DataXu Data. Insight. Action.

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

## Analytics and Visualization
tableau Palantir
OPERA SOLUTIONS metaLayer
METAMARKETS dataspora centrifuge
TERADATA ASTER
SAS TIBCO KARMASPHERE
panopticon Real-Time Visual Data Analysis
Datameer pentaho
platfora ClearStory CIRRO
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

dave@vcdave.com
blogs.forbes.com/davefeinleib

https://blogs-images.forbes.com/davefeinleib/files/2014/06/big-data-landscape-jul-4-2012-00111.png
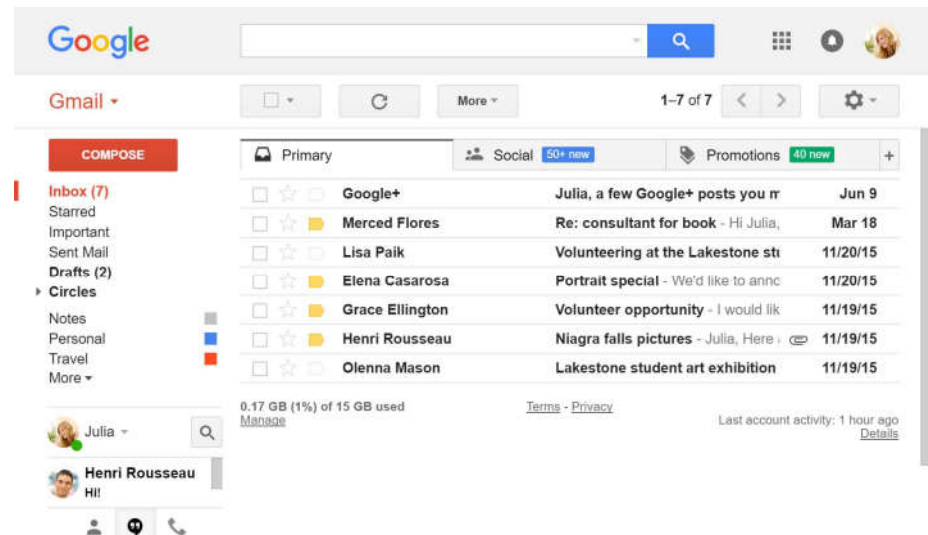
# "The Big Data Application"

# Big Data Technology Application :

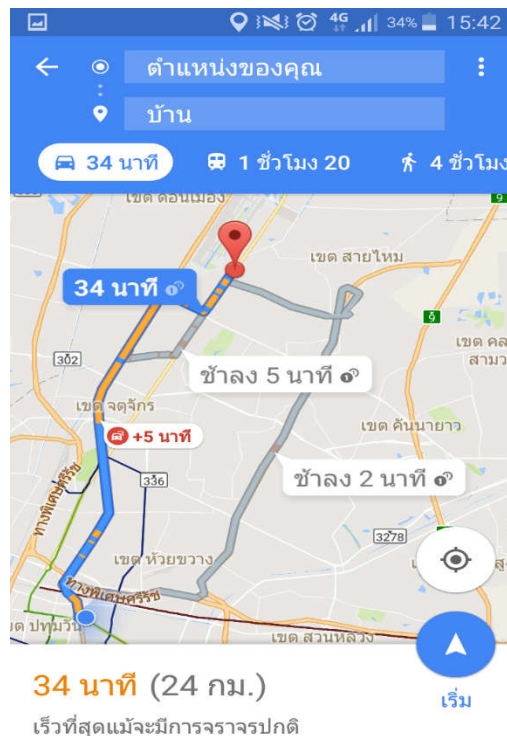**2** Fast Computing

**1** Unlimit Data Storage √
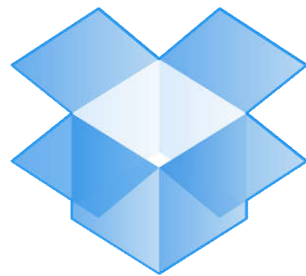
**2** Fast Computing

**3** Service Intelligence
( Data Analytic )

Application
Intelligence



INTRODUCING
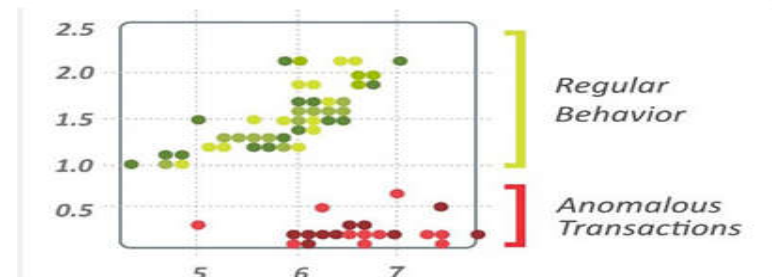amazon echo

Always ready, connected,
and fast. **Just ask.**

**2** Fast Computing

**3** Service Intelligence
( Data Analytic )

Operation
Intelligence



Geolocation

Finding your location: found



Regular
Behavior

Anomalous
Transactions

# What we should to Do :

- Big Data Store

- Big Data Analytic

- Big Data Operation

# Use Case Category:

1. **Service Performance**

2. **Operation Performance**

3. **Social Performance**

4. **Life & Living Performance**

http://www.informationweek.com/big-data/big-data-analytics/5-big-data-use-cases-to-watch/d/d-id/1251031
https://www.qubole.com/resources/solution/best-use-cases-for-big-data-analytics/
https://www-01.ibm.com/software/data/bigdata/use-cases.html

# Use Case Category:

1. **A 360 degree view of the customer**

2. Internet of Things

3. **Data warehouse optimization**

4. Big data service refinery

5. **Information security**

Sentiment Analysis

Customer Micro-Segmentation

Clickstream Analysis

360-Degree Customer View

Ad-hoc Analysis

http://www.informationweek.com/big-data/big-data-analytics/5-big-data-use-cases-to-watch/d/d-id/1251031
https://www.qubole.com/resources/solution/best-use-cases-for-big-data-analytics/
https://www-01.ibm.com/software/data/bigdata/use-cases.html

# Let's set off light bulbs in your head



## Building a Big Data solution

"Building an Effective Data Warehouse Architecture with Hadoop, the cloud, and MPP"

James Serra , Big Data Evangelist , Microsoft

# Data Analytics is needed everywhere

| | | | | | |
|---|---|---|---|---|---|
| Intelligence Gathering | IT infrastructure & Web App optimization | Legal discovery and document archiving | Social network analysis | Traffic flow optimization | Recommenda-tion engines |
| Churn analysis | Location-based tracking & services | Oil & Gas exploration | Weather forecasting for business planning | Healthcare outcomes | Personalized Insurance |
| Fraud detection | Life sciences research | Advertising analysis | Equipment monitoring | Pricing Analysis | Smart meter monitoring |

http://www.zdnet.com/article/getting-big-data-right-is-about-more-than-the-size-of-your-database/

# Personalized Insurance



Insurance companies can help (and some have already started helping) their customers with truly personalized insurance plans tailored to their needs and risks

Personalized policies can reduce costs & better meet customer needs

Insurance Companies can collect real-time data from in-car sensors and combine it with geolocation and in-house systems.  With information such as distance and speed, provide personalized insurance offers based on driving amount, risk, and other factors, for a truly personalized plan that may often save drivers money

$1,600/yr.
US national avg. car insurance premium

# IT Infrastructure and Web App Optimization

IT And Application Managers need to be able understand both immediate and long term metrics to resolve issues and keep costs manageable.

Reduce costs with right-sized infrastructure, and manage issues quickly

A solution that includes HDInsight can manage high-velocity data about server status, health and other metrics - and send alerts when problems occur.

Improve Web Application performance and features by collecting and monitoring usage data and providing access to unstructured data sources in real-time
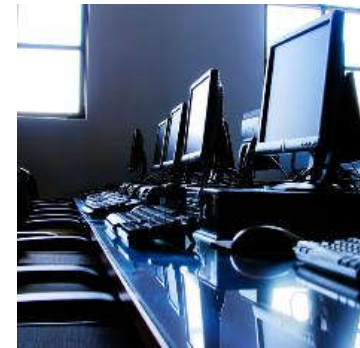
# Recommendation Engines

Retailers can use customer purchase & rating information to serve recommendations to current customers, based on similarities across many dimensions

Significantly improve up-sell and cross-sell opportunities

The vast amount of current and ever-growing customer purchase, rating and click data can all be collected and managed with an Hadoop-based solution, to pinpoint preferences based on purchase history and demographics, and be able to serve useful and compelling cross-sell and up-sell recommendations.

158
Items sold/second by Amazon.com on 11/29/2010 (Cyber Monday)

# Pricing Analysis

Retailers can use customer past purchase, preference, and demo-graphic information to serve real-time custom pricing, instant discounts when near the store.

Significantly improve sales and customer satisfaction

Retailers – whether large, small, online or in-store – can improve margins with more detailed pricing analysis. When a customer is in range of a transaction (either in the store, online or perhaps passing by), offer personalized offers, real-time price quotes, or other frequent-buyer perks to help bring more customers to the store and improve repeat business.

up to 30%
Additional price Mac users accepted for travel from Orbitz

# Advertising Analysis

Marketers can use current page information, past purchase, preference, and demographic information to serve real-time, compelling advertisements that are more likely to be viewed.

Improve return on marketing with improved advertisement response

Improve marketing results by combining public demographic data, browser site history (or past store purchases for store or coupon campaigns), and advertising history into meaningful data analytics that serves relevant advertisements and provides tools for analysis and reporting.

**OTTO**

8%
Click through rate with targeted Hotmail ads

# Customer Churn Analysis



Customers churn happens for a lot of reasons, including quality, service, or feature issues, or new offers from competitors. Individual analysis can help reduce each.

Reduce churn with proactive customer campaigns

To reduce churn, know each customer individually to identify warning signs. With a data analytics solution, demographics and history data can be reviewed and monitored, and proactive efforts can be made to avoid customer churn before it happens.

## 9%
Rate of wireless subscribers switching services in Europe and USA, 2009

# Legal Discovery and Document Archiving

>50%

Of organizations do not track legal hold processes (US, 2012)

Manage documents and content with a data warehouse & analytics solution to find the right content based on searches, semantics analysis and pattern matching

Find docu-ments more quickly; don't miss needed information

Legal cases may necessitate management of a great number of documents that must be identified, collected, stored, processed and reviewed, then turned over to opposing counsel

Large organizations and governments collect a vast number of documents that need to be shared internally or publicly. These need to be organized, searchable, and periodically reviewed
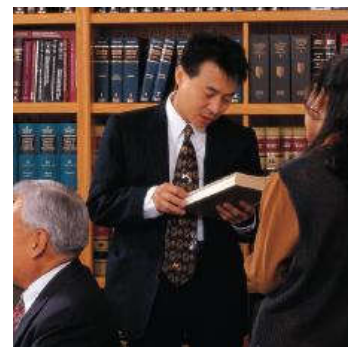
# Retail
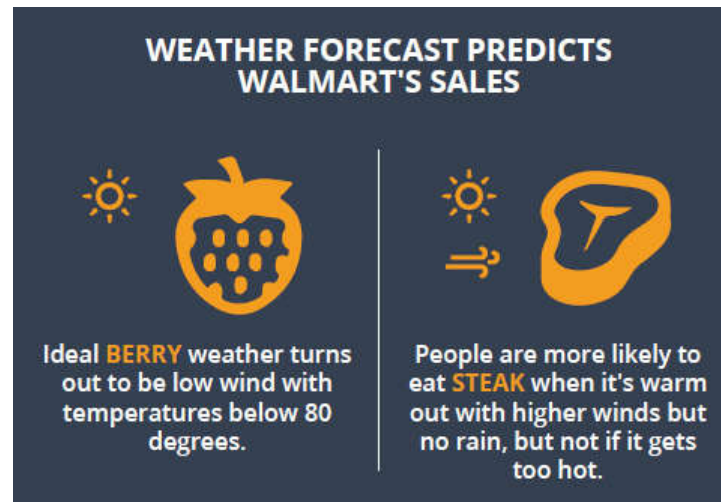
# Using Big data to complete the picture



**1** Social media:
customer sentiment

**2** Bike sensors:
complete journey

**3** Bus GPS:
React to traffic

**4** Wi-Fi:
customer
movement
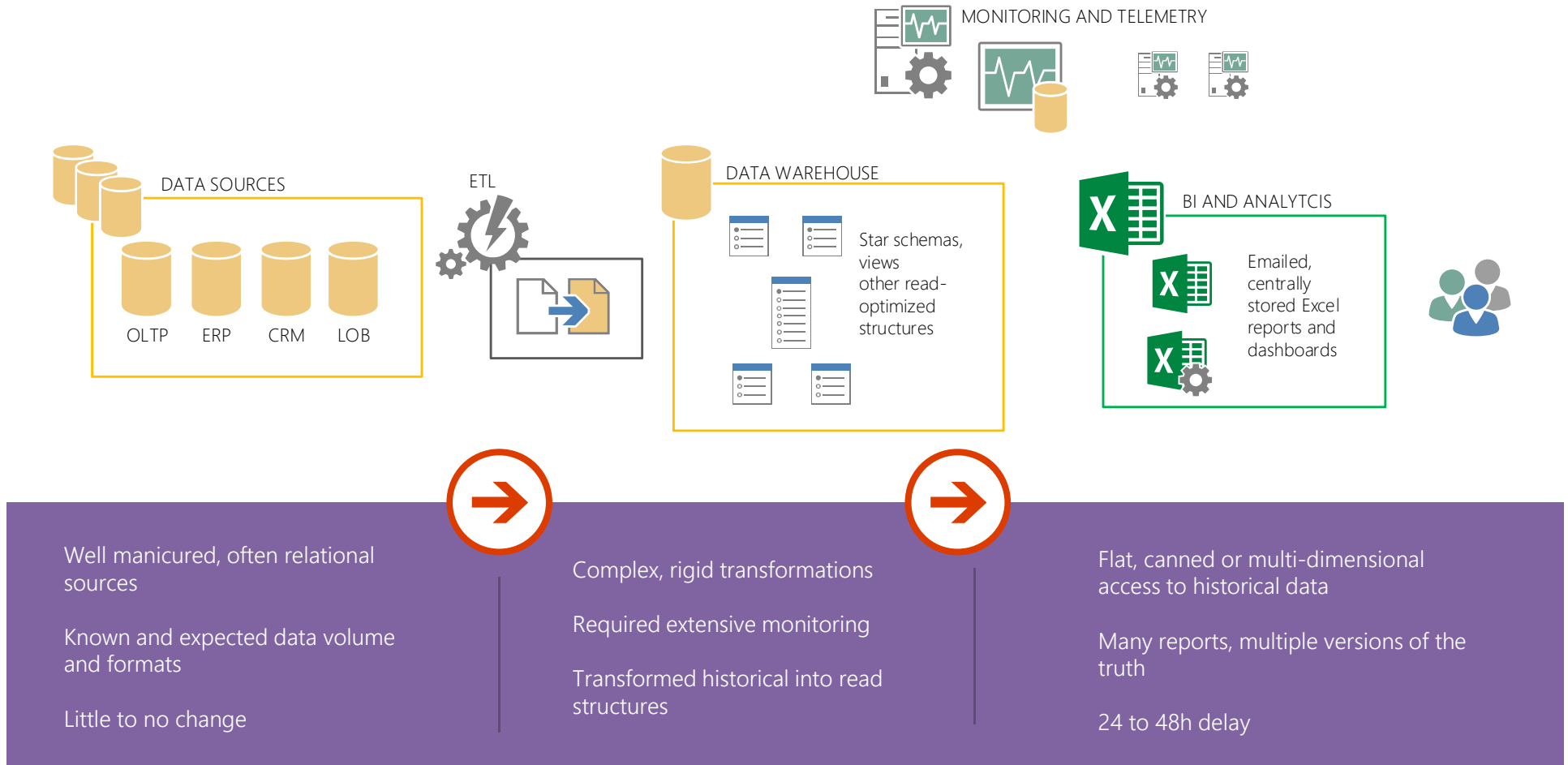in stations

# Modern Data Management

# Data Lake

# What is a data lake?

A storage repository, usually Hadoop, that holds a vast amount of raw data in its native format until it is needed.

- **A place to store unlimited amounts of data in any format inexpensively**
- Allows collection of data that you may or may not use later: "just in case"
- A way to describe any large data pool in which the schema and data requirements are not defined until the data is queried: "just in time" or **"schema on read"**
- Complements EDW and can be seen as a data source for the EDW – capturing all data but only passing relevant data to the EDW
- Frees up expensive EDW resources (storage and processing), especially for data refinement
- **Allows for data exploration to be performed without waiting for the EDW team to model and load the data**
- Some processing in better done on Hadoop than ETL tools like SSIS
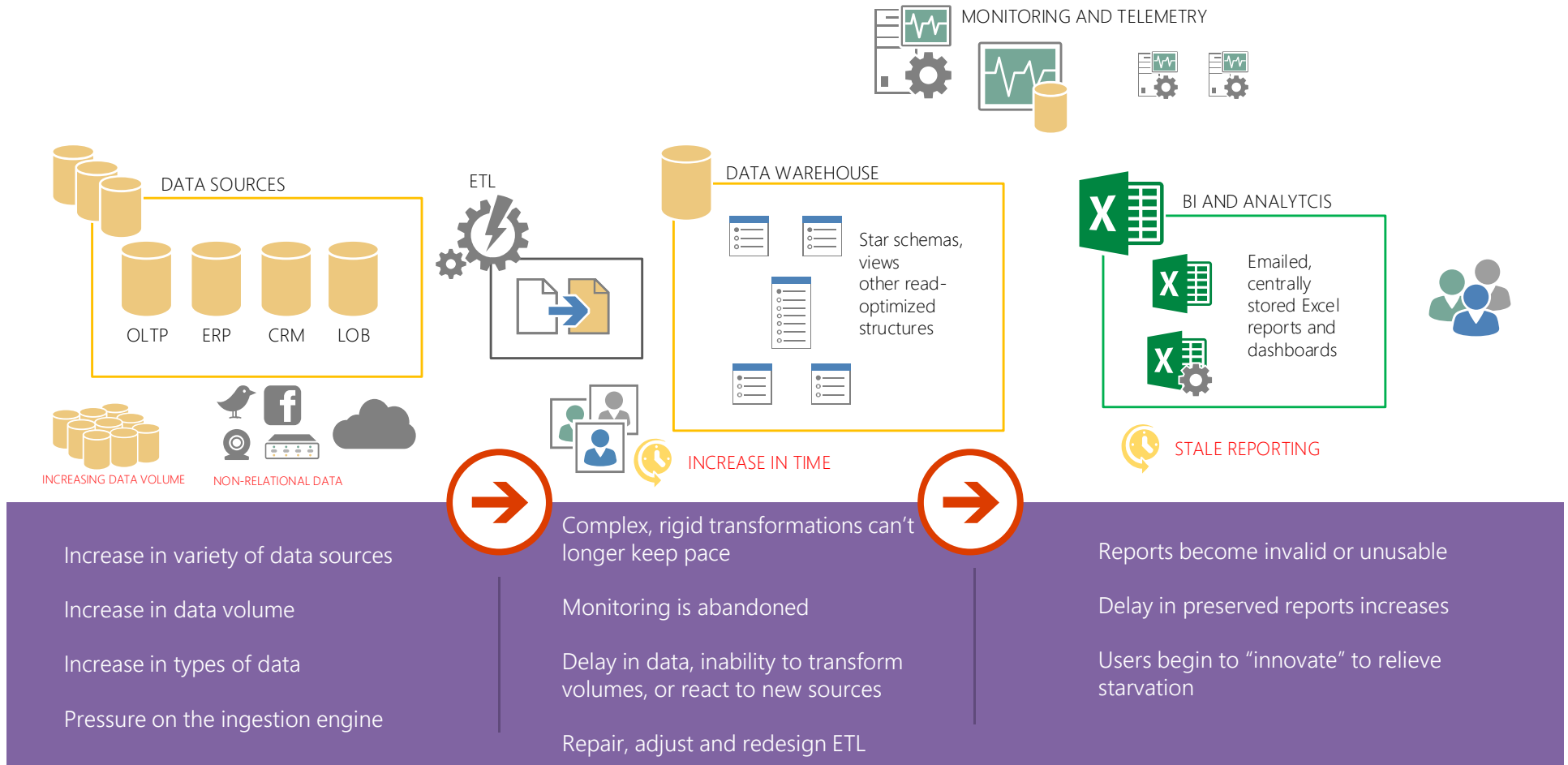- Also called bit bucket, staging area, landing zone or enterprise data hub (Cloudera)

# Traditional Approaches

## Current state of a data warehouse

MONITORING AND TELEMETRY

DATA SOURCES

OLTP  ERP  CRM  LOB

ETL

DATA WAREHOUSE

Star schemas, views other read-optimized structures

BI AND ANALYTCIS

Emailed, centrally stored Excel reports and dashboards

Well manicured, often relational sources

Known and expected data volume and formats

Little to no change

Complex, rigid transformations

Required extensive monitoring

Transformed historical into read structures

Flat, canned or multi-dimensional access to historical data

Many reports, multiple versions of the truth

24 to 48h delay

# Traditional Approaches

## Current state of a data warehouse

MONITORING AND TELEMETRY

DATA SOURCES

ETL

DATA WAREHOUSE

Star schemas, views other read-optimized structures

BI AND ANALYTCIS

Emailed, centrally stored Excel reports and dashboards

OLTP  ERP  CRM  LOB

INCREASING DATA VOLUME

NON-RELATIONAL DATA

INCREASE IN TIME

STALE REPORTING

Increase in variety of data sources

Increase in data volume

Increase in types of data

Pressure on the ingestion engine

Complex, rigid transformations can't longer keep pace

Monitoring is abandoned

Delay in data, inability to transform volumes, or react to new sources

Repair, adjust and redesign ETL

Reports become invalid or unusable

Delay in preserved reports increases

Users begin to "innovate" to relieve starvation

# Traditional Approaches

## Impact if we keep the current architecture

- Drop useful data by introducing ETL "bias"

- Potentially insightful data is lost

- Create latency as volumes increase and sources change

- Duplicate data through staging environments to support ETL

- Expensive "reactive" hardware to support processing scale requirements

# New Approaches

## Data Lake Transformation (ELT not ETL)

**DATA SOURCES**

OLTP  ERP  CRM  LOB

NON-RELATIONAL DATA

FUTURE DATA SOURCES

**EXTRACT AND LOAD**

**DATA LAKE**

**DATA WAREHOUSE**

Star schemas, views other read-optimized structures

**BI AND ANALYTCIS**

Discover and consume predictive analytics, data sets and other reports

**DATA REFINERY PROCESS (TRANSFORM ON READ)**

Transform relevant data into data sets

**OTHER REFINERY PROCESSES**
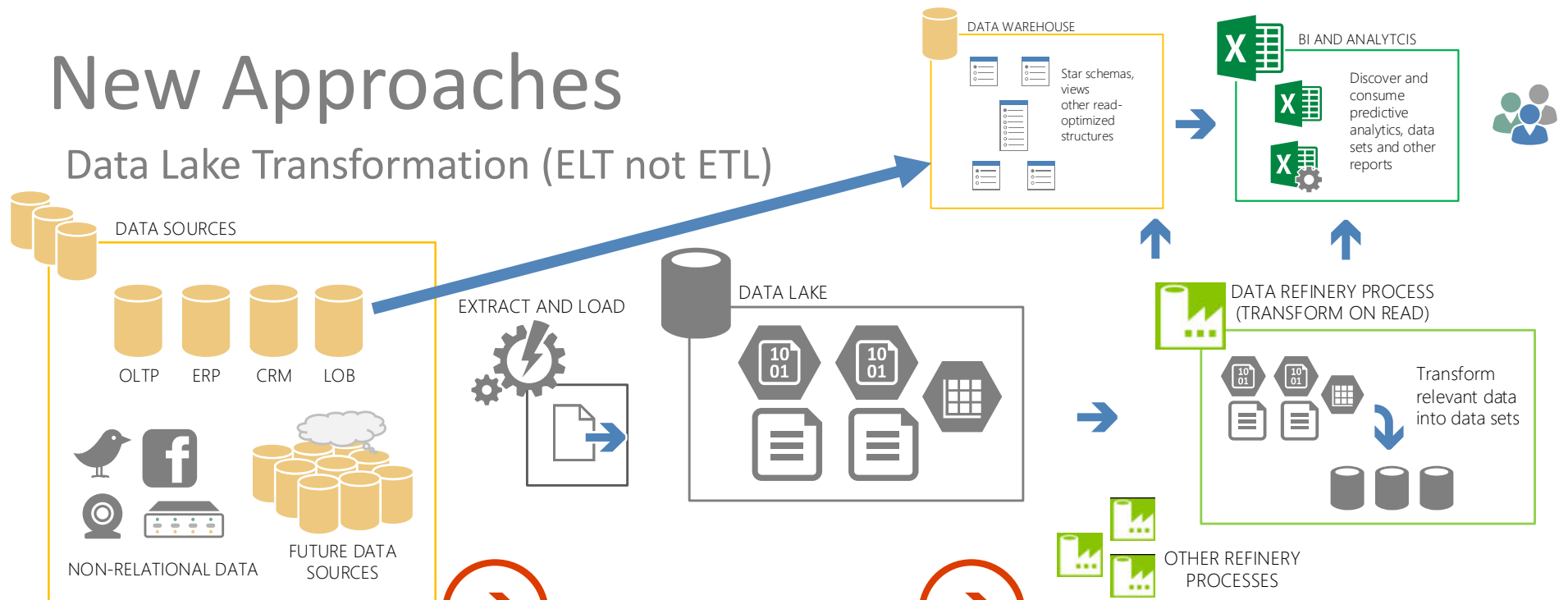
All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data

Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

Streaming data accommodation becomes possible

Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

Users discover published data sets/services using familiar tools

# New Approaches

## Change architecture for Analyst's Needs

- Entire "universe" of data is captured and maintained

- Mining of data via transformation on read leaves all data in place

- Refineries leverage the power of the cloud and traditional technologies

- Integration with traditional data warehousing methodologies

- Scale can be pushed to cloud for more horsepower

- Orchestration of data is a reality (less rigid, more flexible, operational)

- Democratization of predictive analytics, data sets, services and reports

# Use cases using Hadoop and a DW in combination

### Bringing islands of Hadoop data together
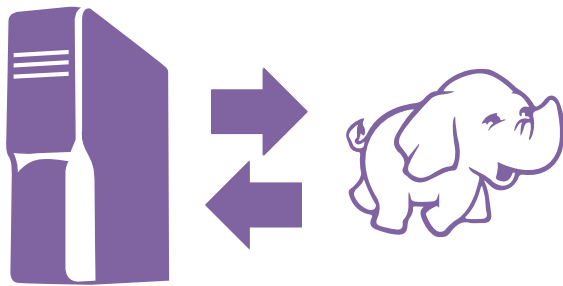
Archiving data warehouse data to Hadoop (move)
(Hadoop as cold storage)

Exporting relational data to Hadoop (copy)
(Hadoop as backup/DR, analysis, cloud use)
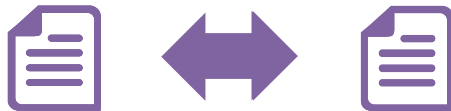
Importing Hadoop data into data warehouse (copy)
(Hadoop as staging area, sandbox, Data Lake)

# Modern Data Warehouse

# Modern Data Warehouse

Think about future needs:

- Increasing data volumes
- Real-time performance
- New data sources and types
- Cloud-born data
- Multi-platform solution
- Hybrid architecture

# Modern data warehouse defined

## BI & ANALYTICS

| Self-service | Corporate | Collaboration | Mobile | Predictive |
|---|---|---|---|---|

## DATA ENRICHMENT AND FEDERATED QUERY

| Single query model | Extract, transform, load | Data quality | Master data management |
|---|---|---|---|

## DATA MANAGEMENT & PROCESSING

| Relational | Non-relational | Analytical | Streaming | Internal & External |
|---|---|---|---|---|

## INFRASTRUCTURE

### Data sources
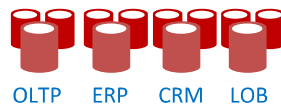
OLTP    ERP    CRM    LOB

### Non-relational data

Devices    Web    Sensors    Social

# Federated Querying

# Federated Querying

Other names: Data virtualization, logical data warehouse, data federation, virtual database, and decentralized data warehouse.

*A model that allows a single query to retrieve and combine data as it sits from multiple data sources, so as to not need to use ETL or learn more than one retrieval technology*
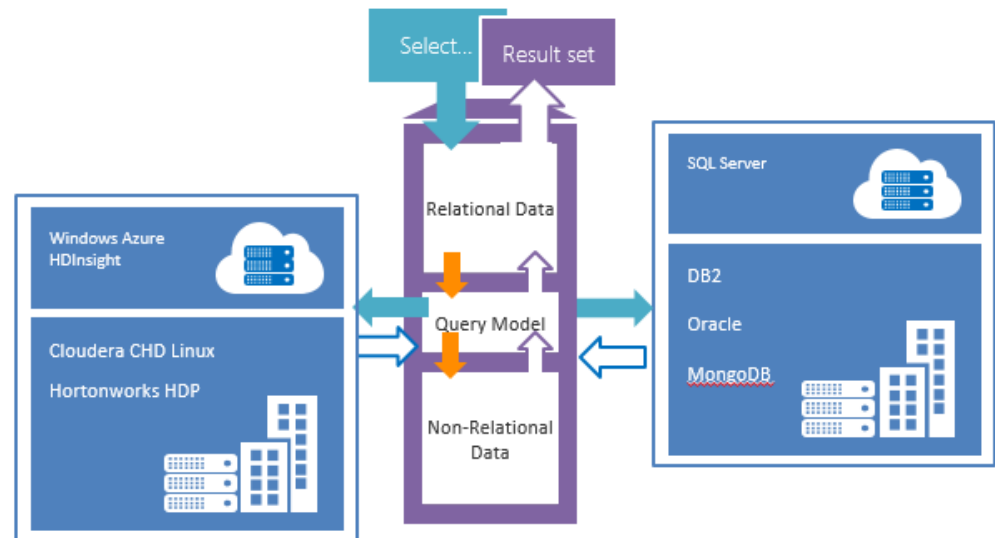
# Do you have any of these pain points?

Are you using or going to use "Big Data" and/or "Hadoop"

No or limited access to detailed data; can only surface reports and cannot ask ad-hoc questions.

Slow data loading performance cannot keep up with the need for data from transactional systems for intraday reporting.

MOLAP cube processing and data refresh take too long.

Slow query performance with need for constant tuning, especially with SAN storage.
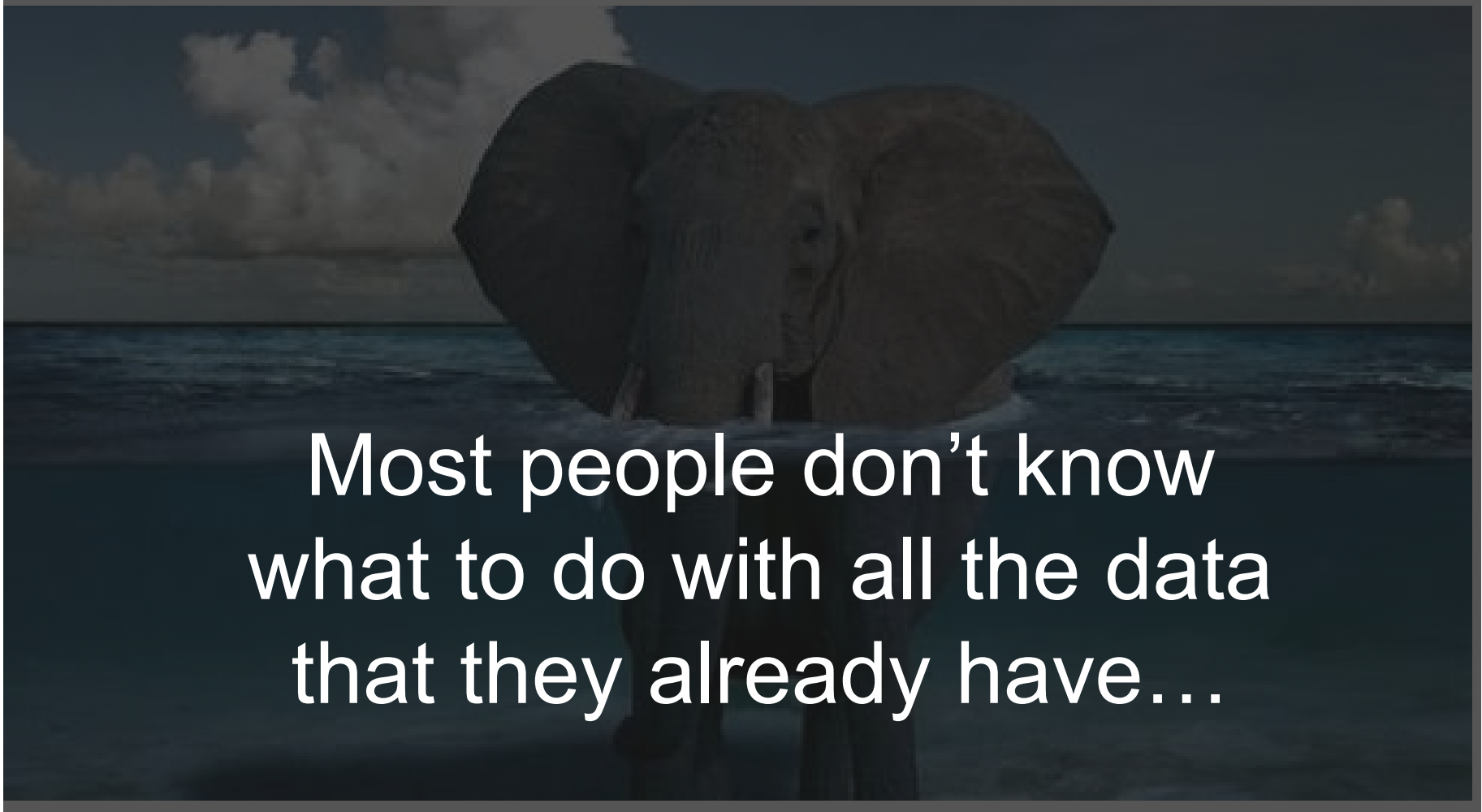
High cost of SAN storage chargeback.

# How
# can we take Advantage from
## Big Data Technology
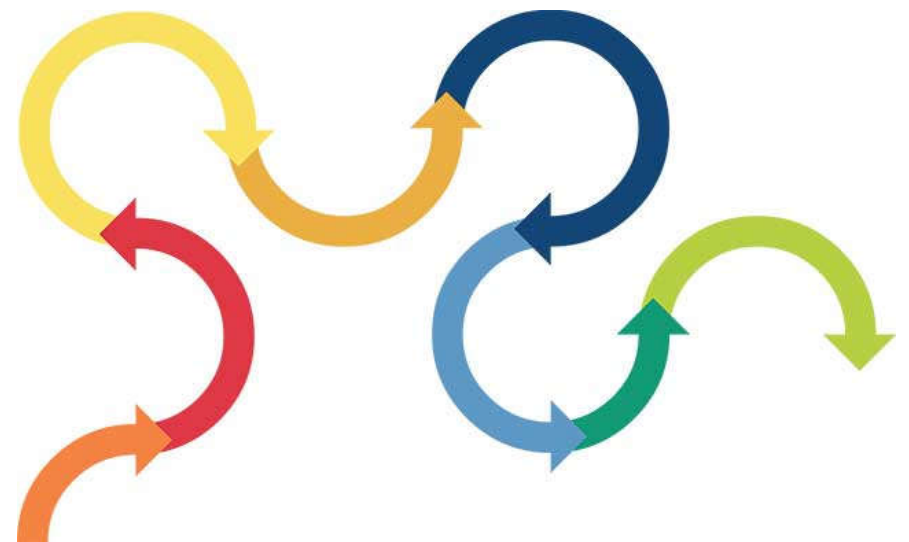
Big Data isn't *big*,
if you *know* how to
*use it*.

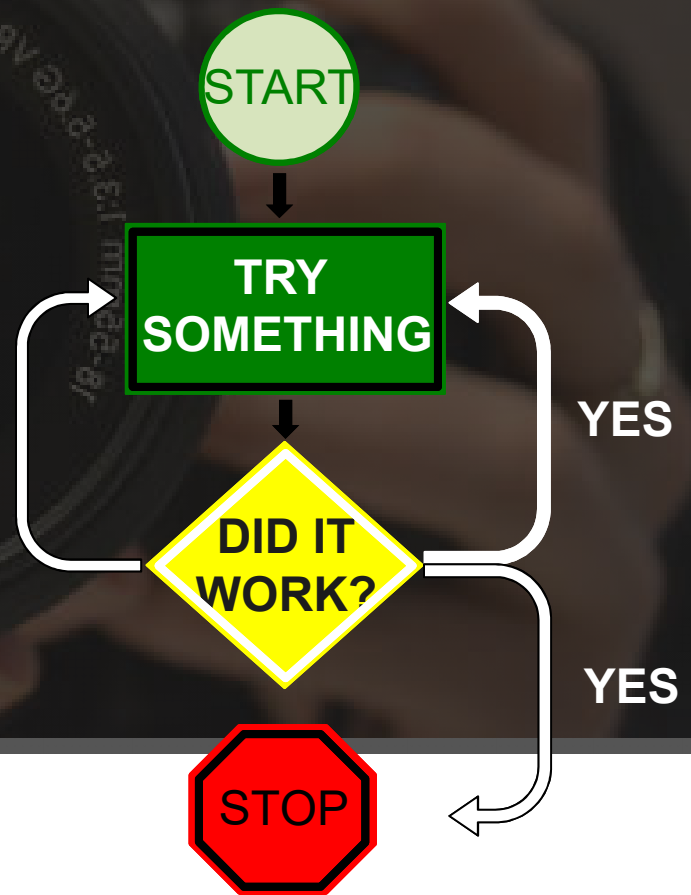Most people don't know
what to do with all the data
that they already have…

# Get Big

by starting

*small*

Focus on *Business Impact.*

# Introduction to Big Data

## Q & A

# Resources

- The Modern Data Warehouse: http://bit.ly/1xuX4Py
- Fast Track Data Warehouse Reference Architecture for SQL Server 2014: http://bit.ly/1xuX9m6
- Should you move your data to the cloud? http://bit.ly/1xuXbKU
- Presentation slides for Modern Data Warehousing: http://bit.ly/1xuXcP5
- Presentation slides for Building an Effective Data Warehouse Architecture: http://bit.ly/1xuXeX4
- Hadoop and Data Warehouses: http://bit.ly/1xuXfu9
- What is the Microsoft Analytics Platform System (APS)? http://bit.ly/1xuXipO
- Parallel Data Warehouse (PDW) benefits made simple: http://bit.ly/1xuXlSy
- What is Advanced Analytics? http://bit.ly/1LDklkB

# Activity Data

Simple activities like listening to music or reading a book are now generating data. Digital music players and eBooks collect data on our activities. Your smart phone collects data on how you use it and your web browser collects information on what you are searching for. Your credit card company collects data on where you shop and your shop collects data on what you buy. It is hard to imagine any activity that does not generate data.

# Conversation Data

Our conversations are now digitally recorded. It all started with emails but nowadays most of our conversations leave a digital trail. Just think of all the conversations we have on social media sites like Facebook or Twitter. Even many of our phone conversations are now digitally recorded.

# Photo and Video Image Data

Just think about all the pictures we take on our smart phones or digital cameras. We upload and share 100s of thousands of them on social media sites every second. The increasing amounts of CCTV cameras take video images and we up-load hundreds of hours of video images to YouTube and other sites every minute .

# Sensor Data

We are increasingly surrounded by sensors that collect and share data. Take your smart phone, it contains a global positioning sensor to track exactly where you are every second of the day,  it includes an accelometer to track the speed and direction at which you are travelling. We now have sensors in many devices and products.

# The Internet of Things Data

We now have smart TVs that are able to collect and process data, we have smart watches, smart fridges, and smart alarms. The Internet of Things, or Internet of Everything connects these devices so that e.g. the traffic sensors on the road send data to your alarm clock which will wake you up earlier than planned because the blocked road means you have to leave earlier to make your 9am meeting…

.