


# BIG-DATA ANALYTICS

## ~~A PRIMER FOR THE ADVENTUROUS~~

### C'MON LETS BE HONEST!



Arijit Mitra, M.Sc.  
Apurba Technologies Inc.  
April 23<sup>rd</sup>, 2016  
ACCU 2016c

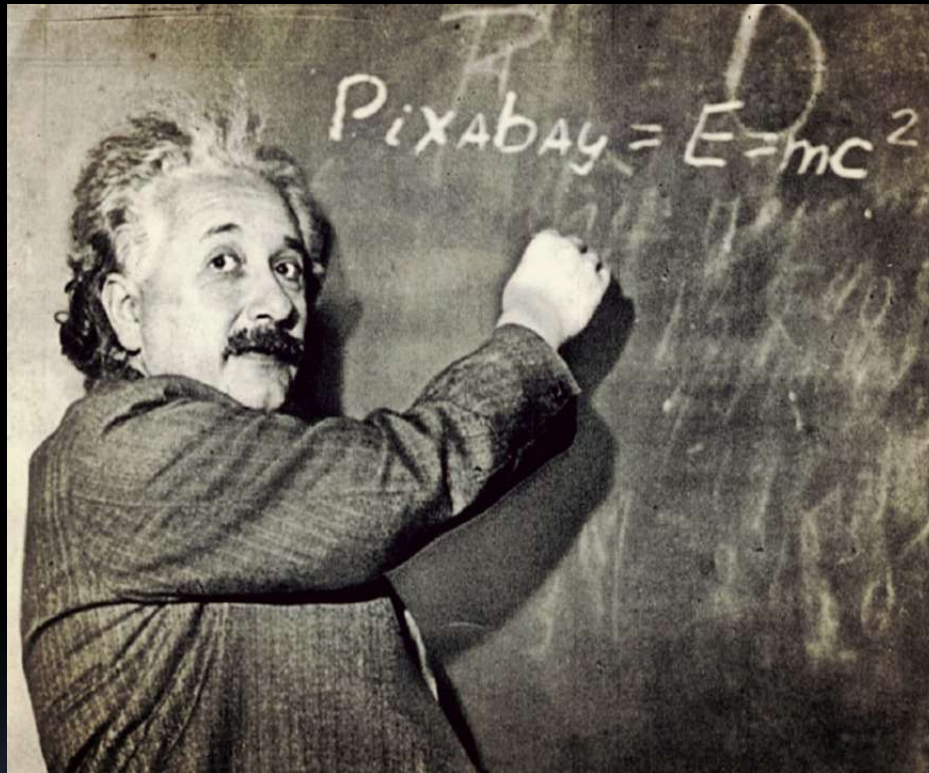
# Section #1: Who am I?

More importantly, what am I doing here?



CTO of a company  
developing Big Data  
Solutions

Why am I here?



To explain  
Big Data –  
Simply!!!

What's in it for us?



Big Data = Big Money

# Objectives?

Big Data Overview

Applying IT



# Practical Examples



# Application in Health Care



# Application in Finance







# Arijit Mitra

- M.Sc. Electronic Engineering
- Worked in diverse industry sectors, from Finance, Legal, Energy, Bio-pharmaceutical and Government
- 16 Years+ Enterprise Content Management
- Founder hi-tech startup, HQ Silicon Valley, focusing on Big Data Analytics

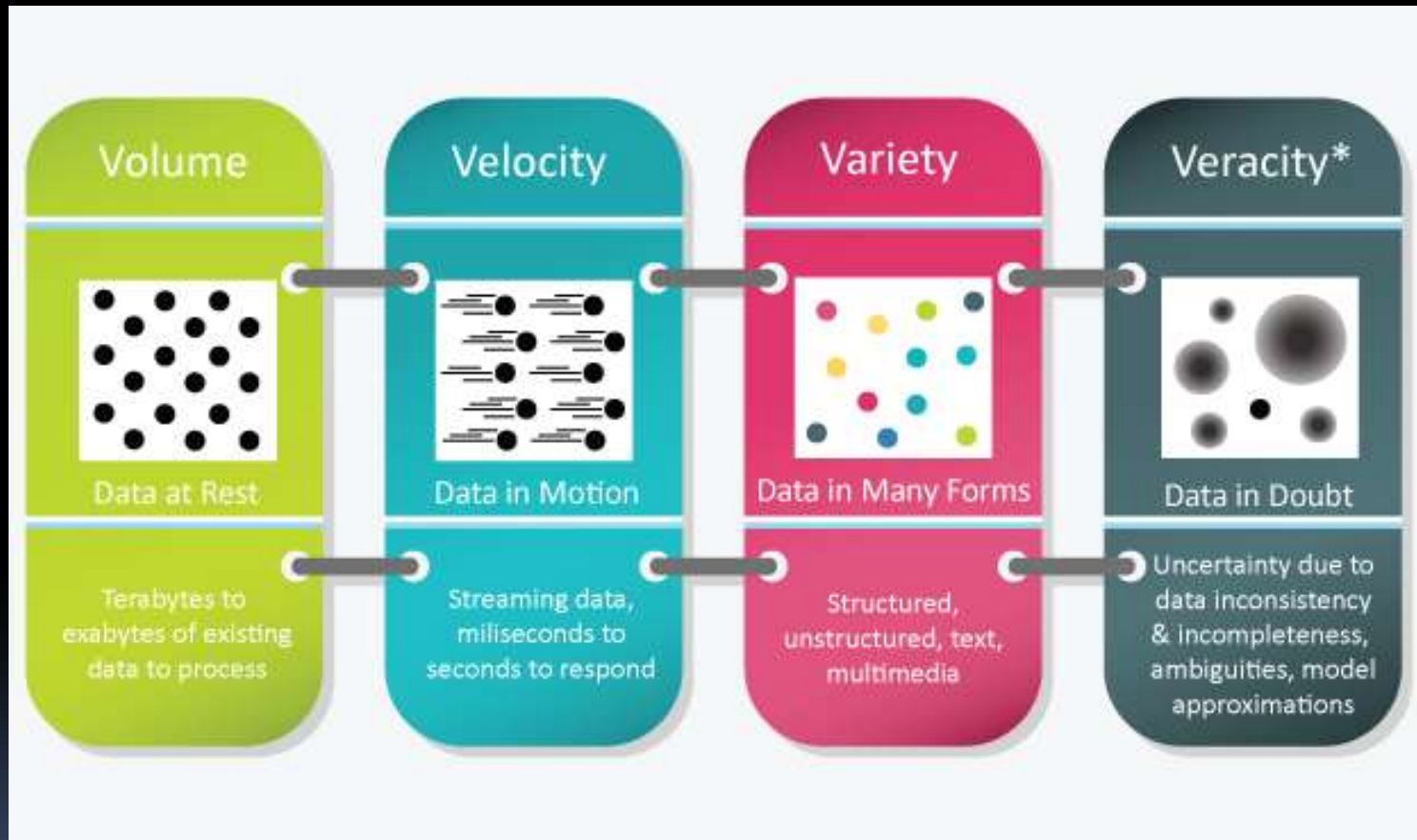
# Section #2: Concepts



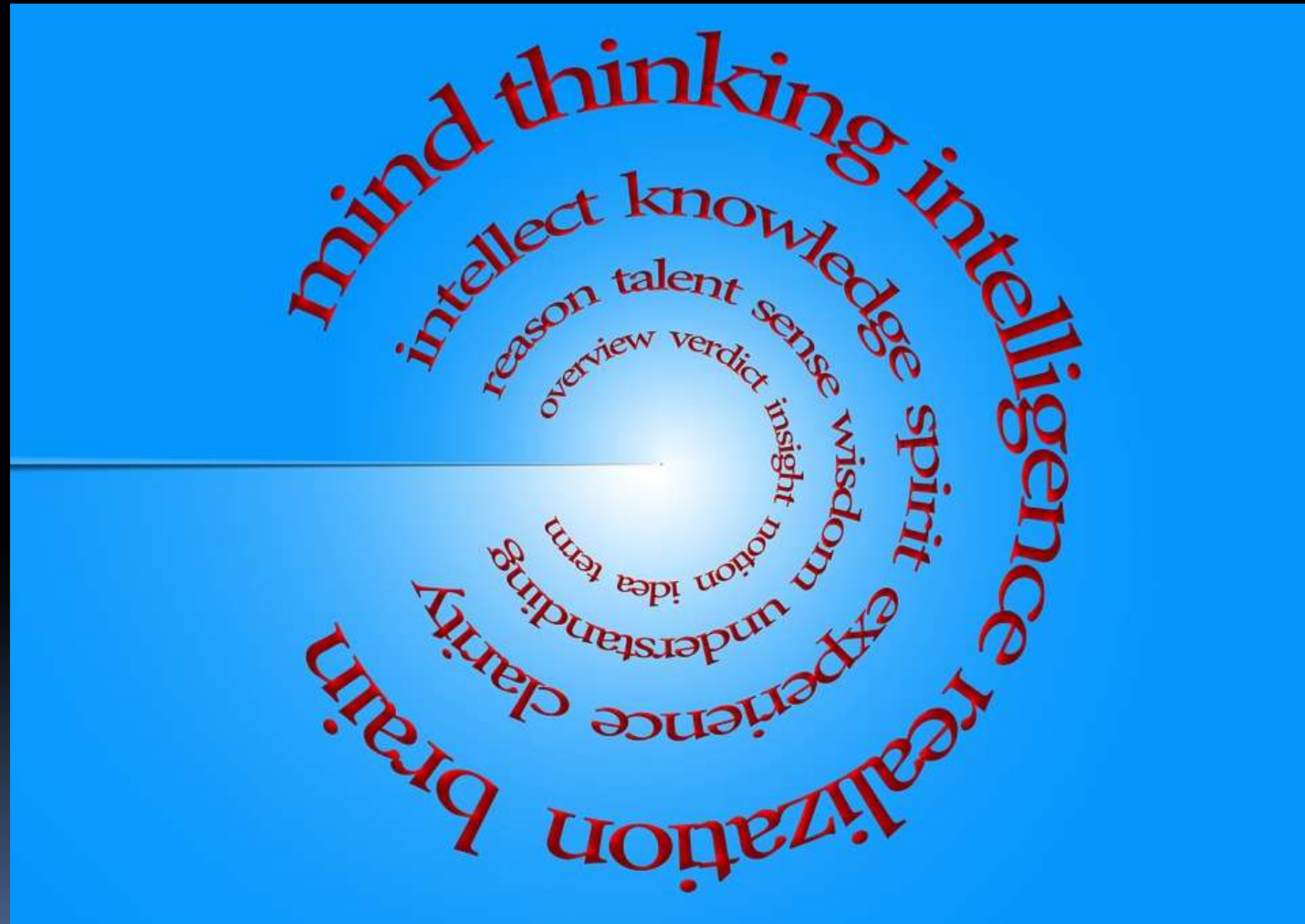
# Gartner Definition – Big Data

- “Big data” is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making
- <http://www.gartner.com/it-glossary/big-data>

# Big Data – States of Matter



# Data promotes Thought



# Volume: Big, Bigger & Bigger

- Gigabyte = 1,000 megabytes
- Terabyte = 1,000 gigabytes
- Petabyte = 1,000 terabytes
- Exabyte = 1,000 petabytes

Velocity: Blink and you'll miss  
it



# Variety: Poly-WTF!



Structured Data



Unstructured Data



# Veracity: Data Truth






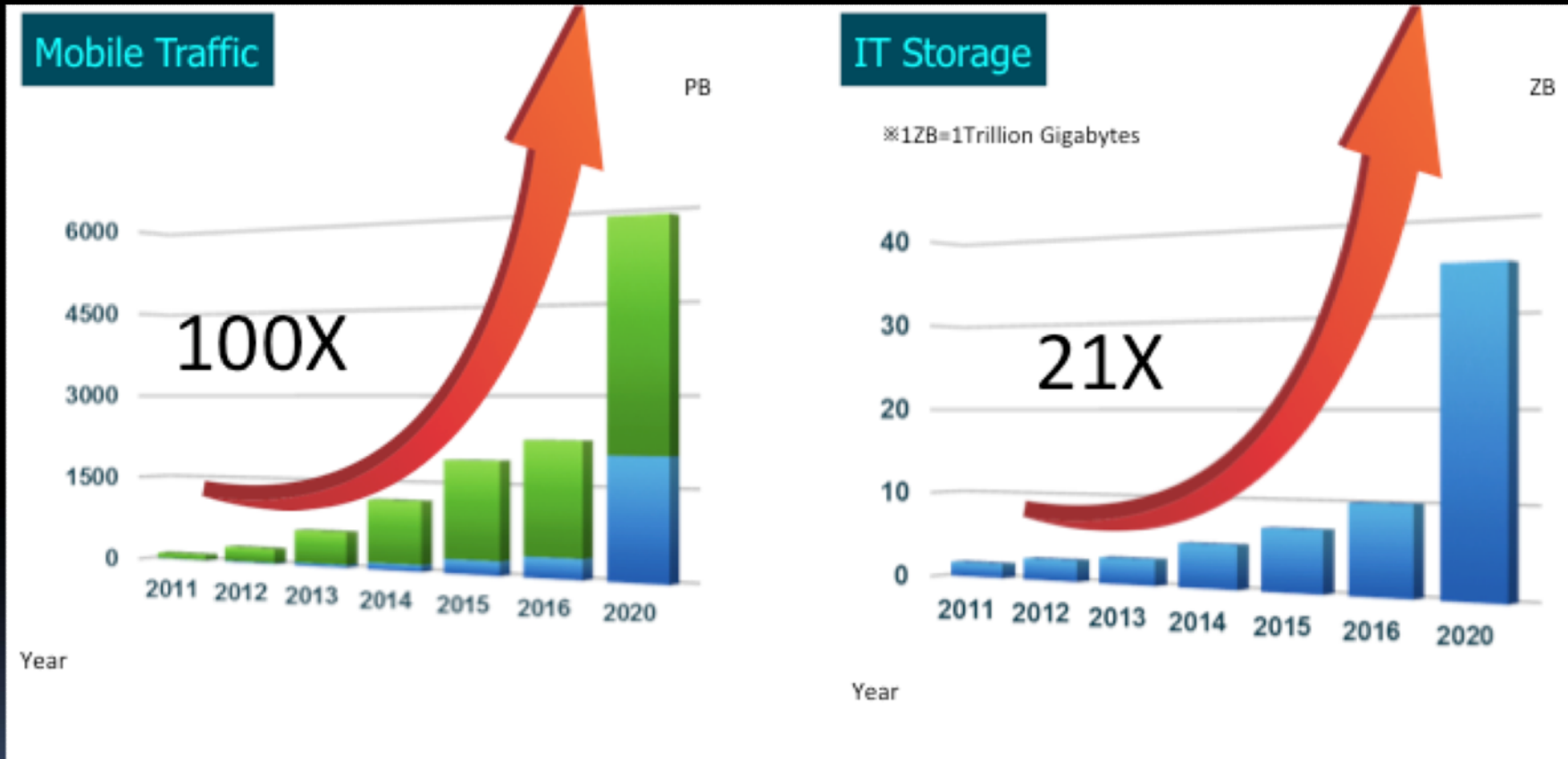
# Section #3: Origins

## A brief history of Big-data

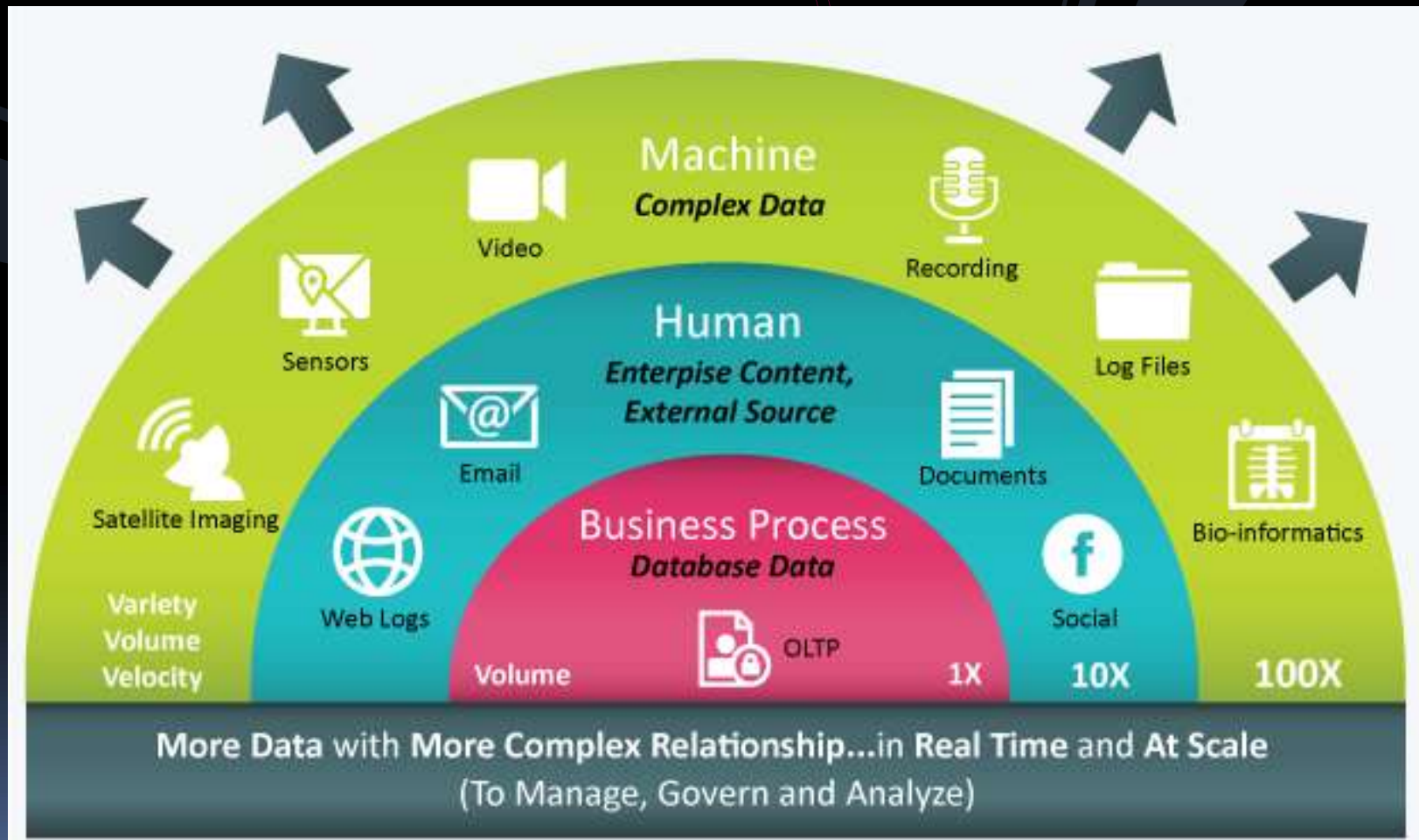
### Rippling across time – evolution of ideas

- 1941 – Phrase Information explosion
  - 1944 – Fremont Rider speculates that the Yale Library would house – 2m volumes, span 6000 shelves, 6000 librarians
  - 1997 – Term 'Big Data' used in a paper by Michael Cox and David Ellsworth
  - 1999 – “Big Data for Scientific Visualization” - Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes
  - 2000 - Peter Lyman and Hal R. Varian at UC Berkeley publish “How Much Information?”
  - 2001 – Terms Volume, Velocity, Variety coined by Doug Laney, an analyst with the Meta Group
- 

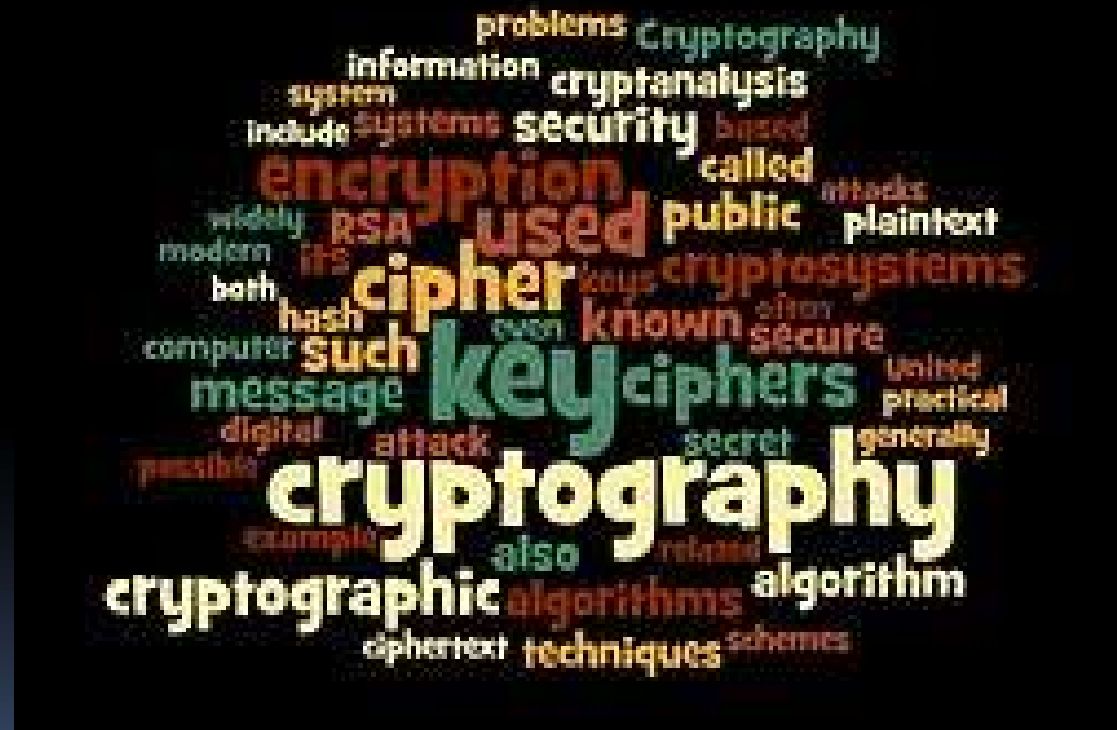
# Ever Expansion



# Where is this data coming from?

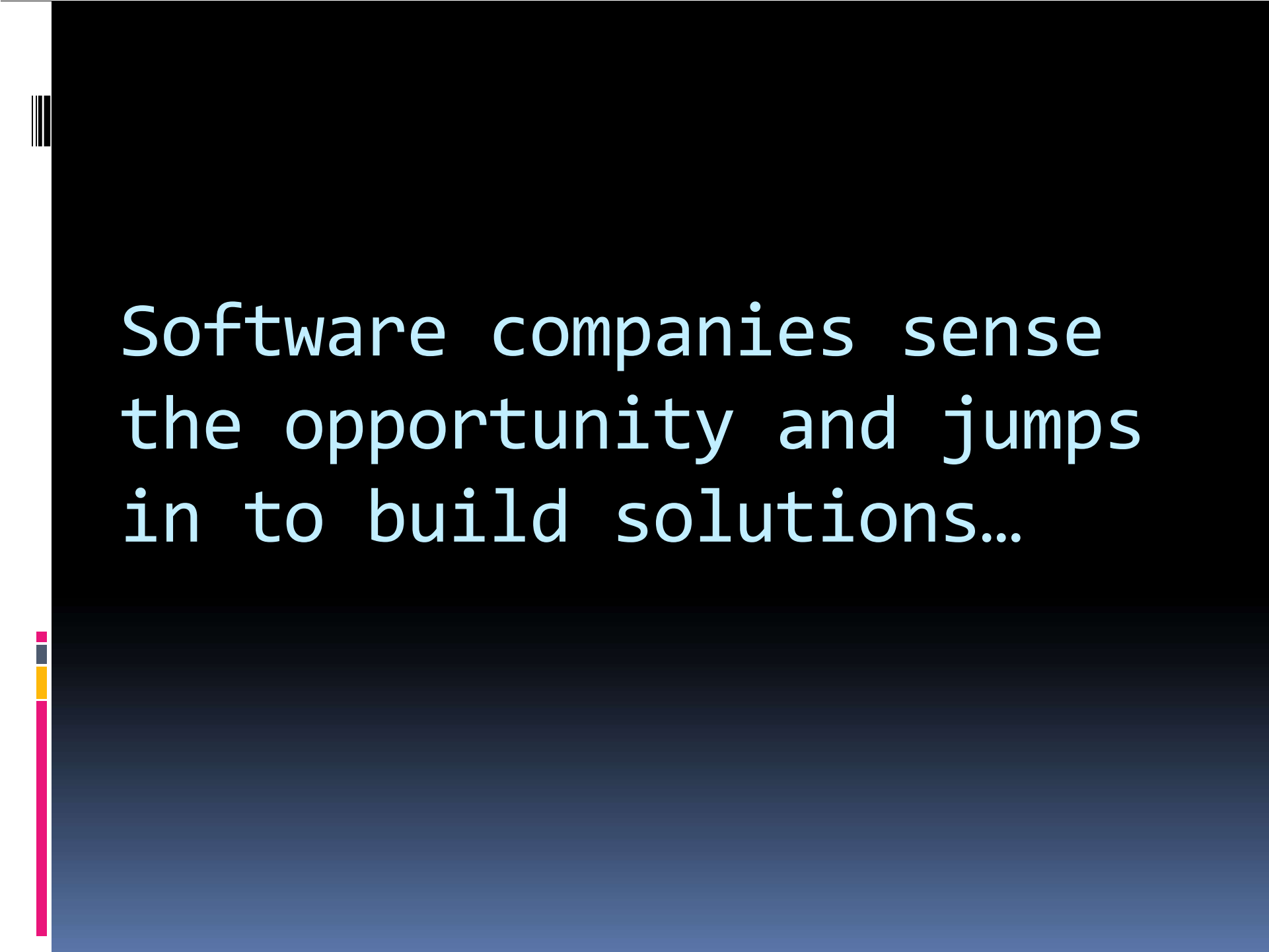


# Privacy? Huh??





# Promise of Data Insight



Software companies sense  
the opportunity and jumps  
in to build solutions..



# Section #4: Why? Who? How? What?



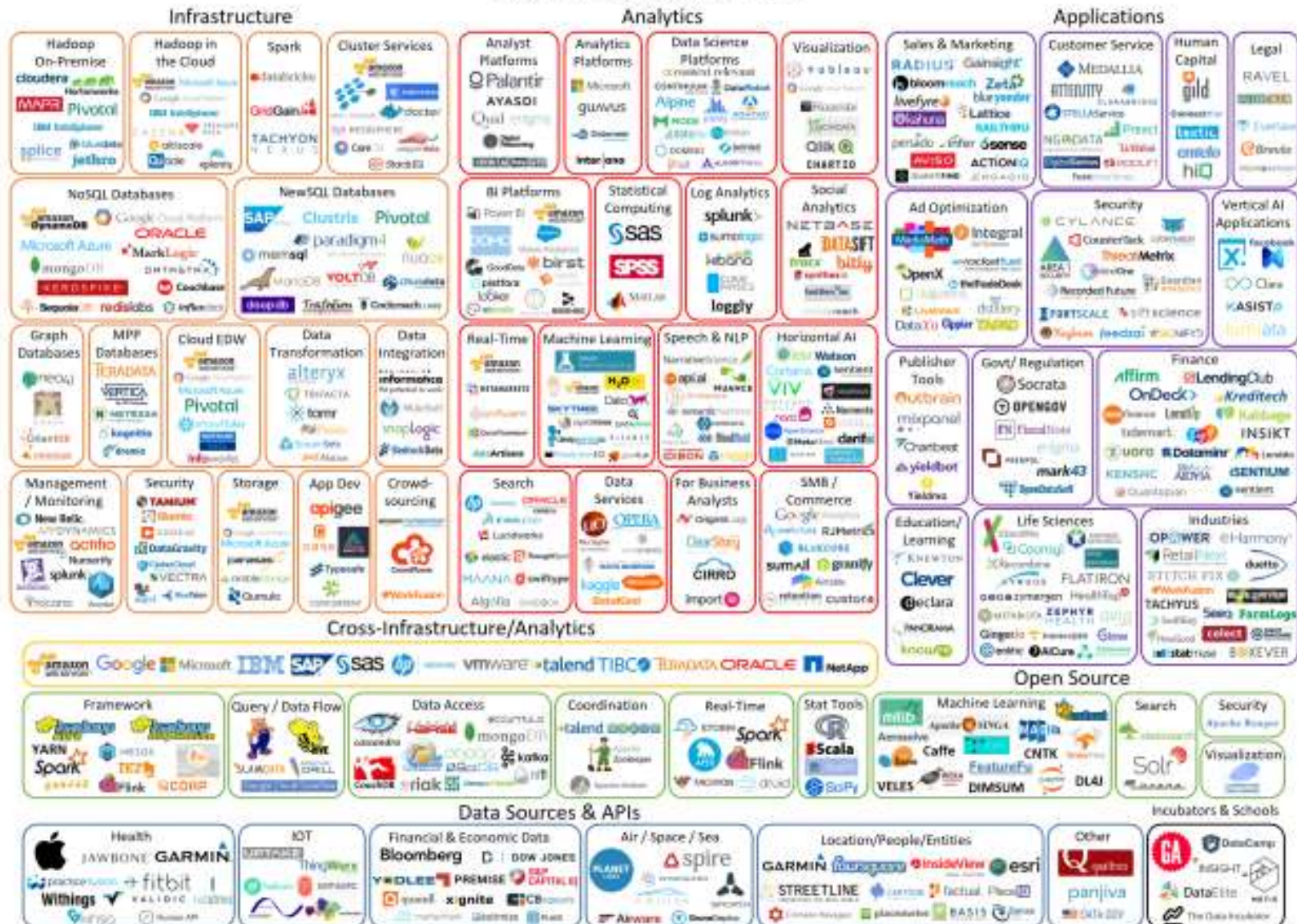
WHY?

# Big Data Science Tackling Humanities Biggest Challenges



# WHO?

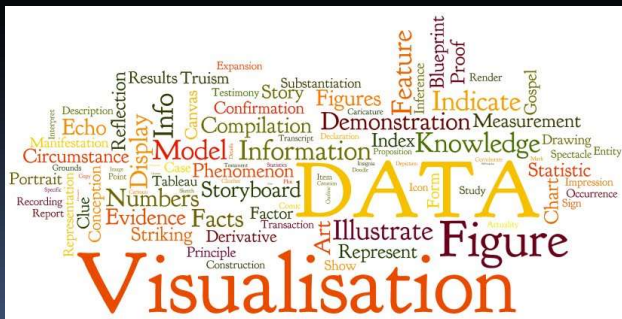
## Big Data Landscape 2016



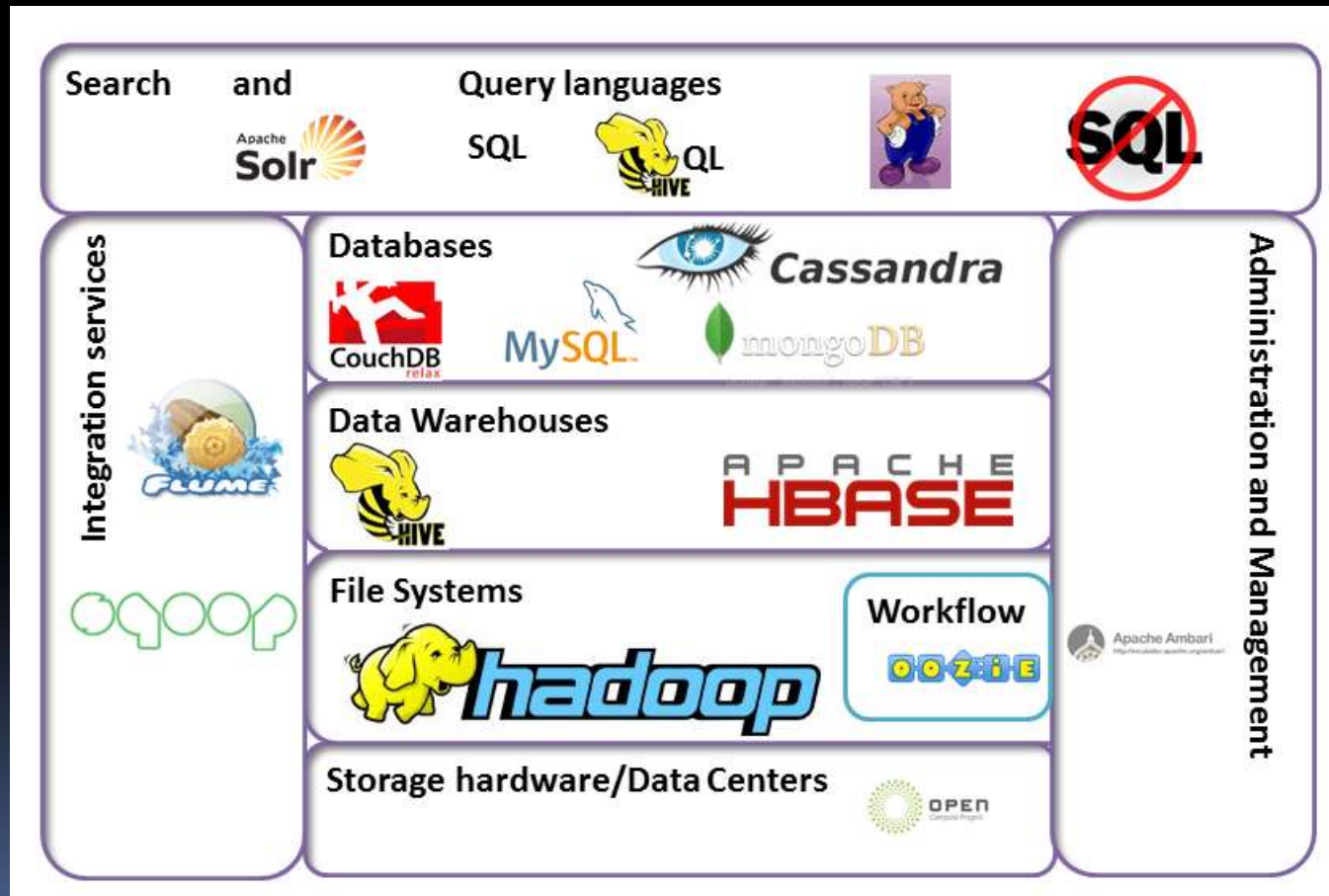
© Matt Turck (@mattturck), Jim Hao (@jimhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

# How?

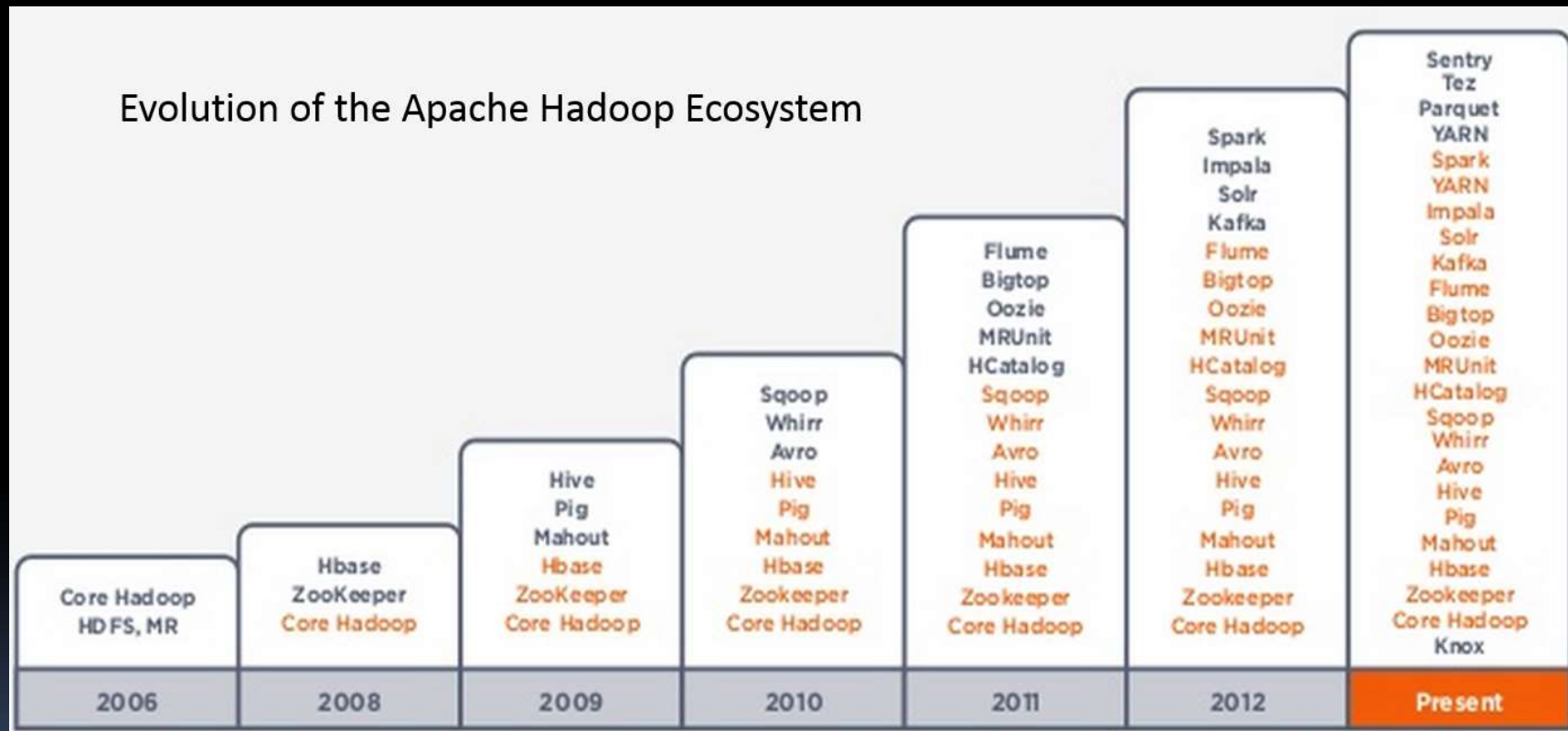


# What?



# Continuous Evolution

Evolution of the Apache Hadoop Ecosystem

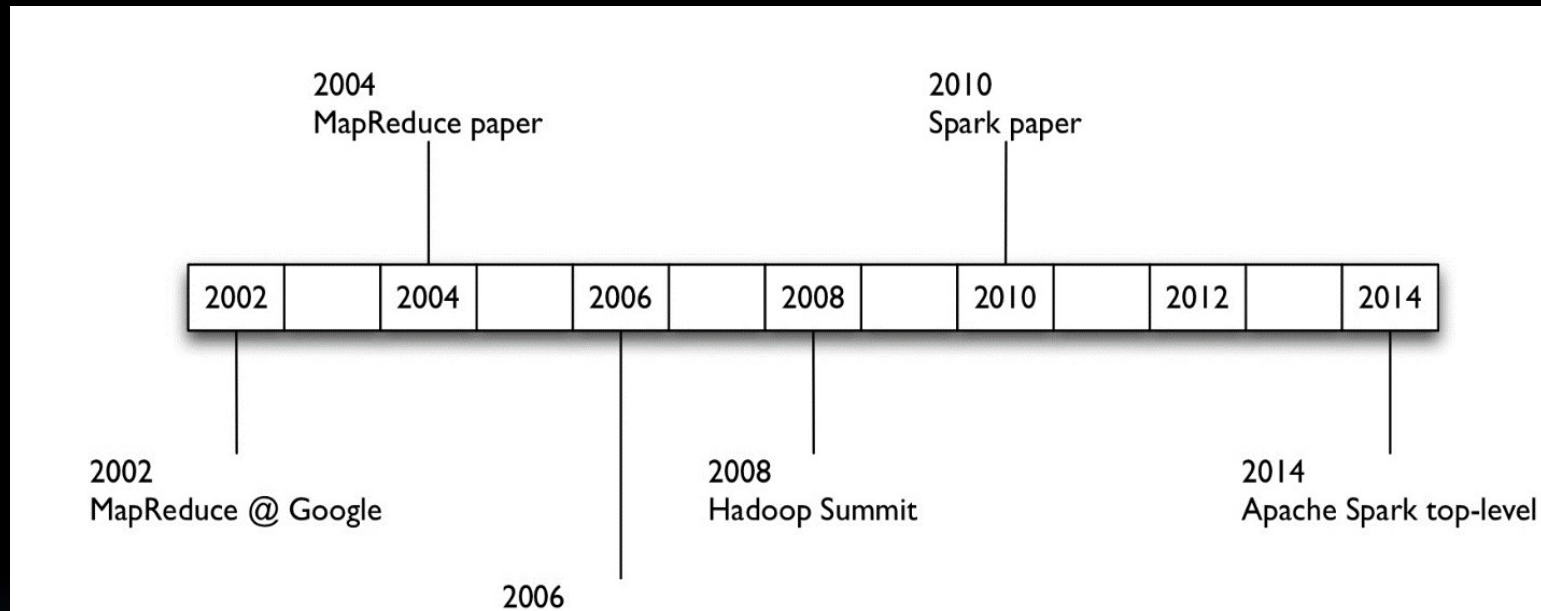


# Section #5: Technology





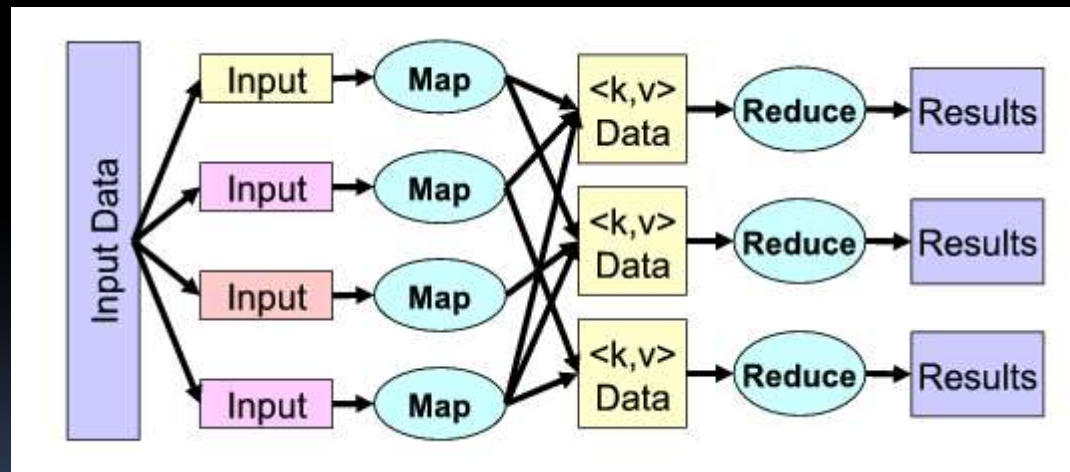
# Computational framework





# MapReduce

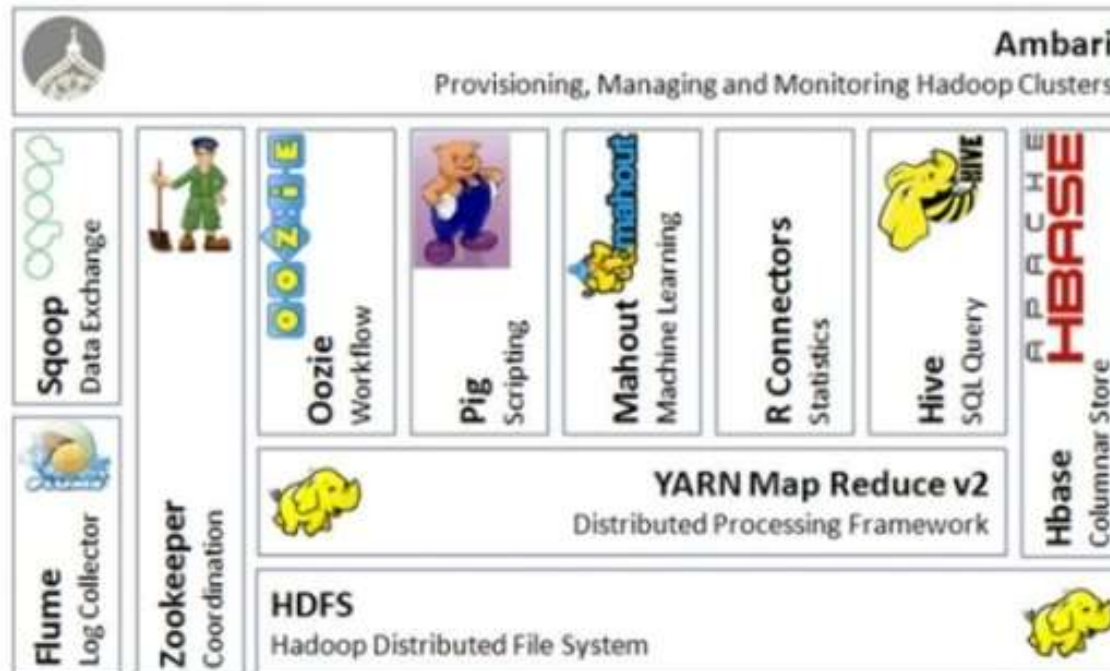
- In 2004, Google published a paper on a process called MapReduce that used such an architecture.
- 2005 Apache Open Project Hadoop adopts MapReduce



# Hadoop project



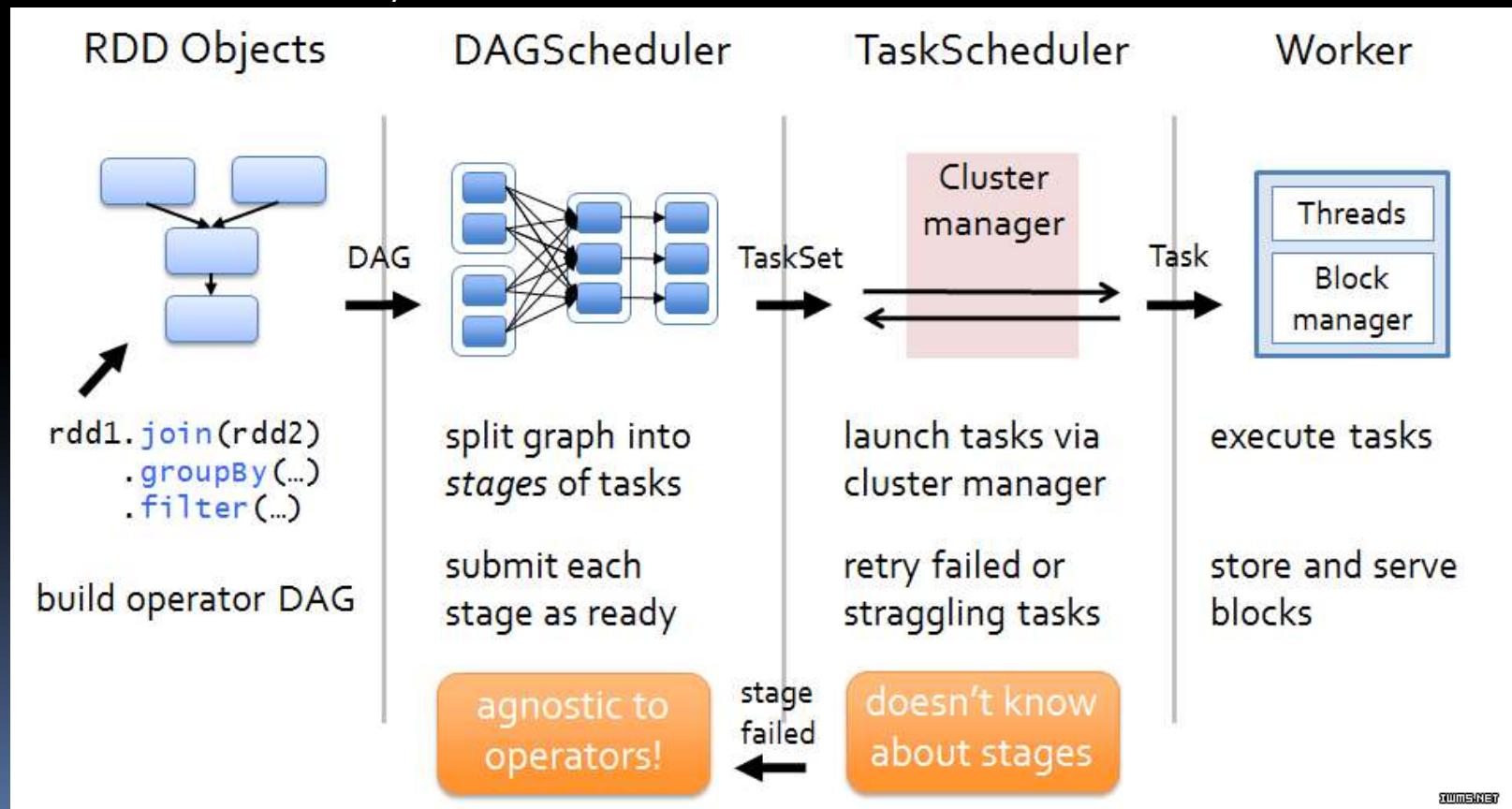
## Hadoop Ecosystem



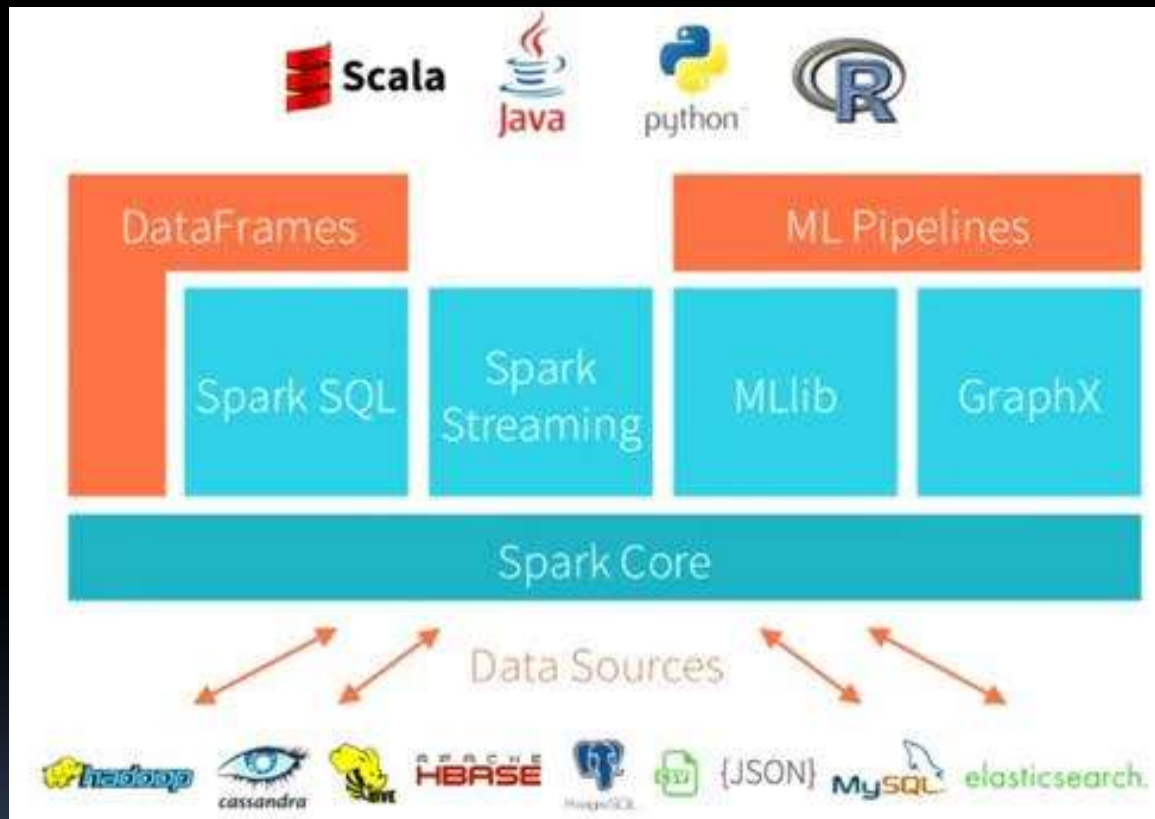
Note: This is not an exhaustive list

# Apache Spark's Resilient Distributed Data (RDD)

- RDD concept & implementation first described: Resilient Distributed Datasets: A Fault Tolerant Abstraction for In-Memory Cluster Computing – University of California- Berkeley.




# Apache Spark





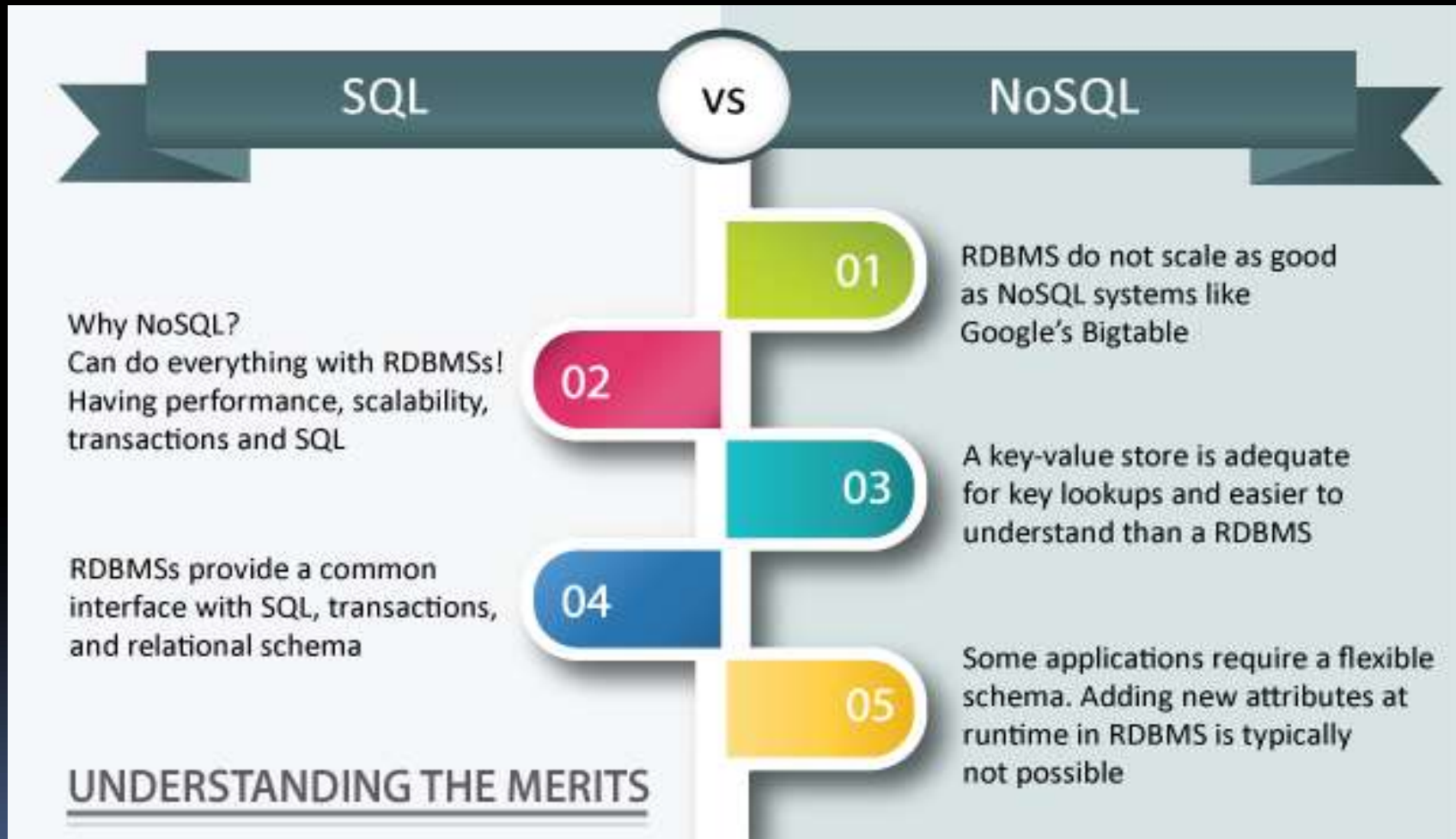
# Apache Spark vs Hadoop MapReduce

- We found Apache Spark
    - Easier to Install
    - Much Faster
    - More Flexible
    - Readily Scalable
    - Easier to code
- 

# New Age-Old Questions



# SQL or NoSQL





## Section #6: Using It!

How is this really  
implemented?





# Coders View

- Download JDK 1.8.0\_45
- Download [Eclipse](#)
- Download [DataStax Cassandra](#)
- Download [Apache Spark](#)
  - select package type from 2nd Drop Down: Pre Built for Hadoop 2.6 and later
  - Download hadoop-common-2.2.0-bin-master.zip
  - Copy to C:\ drive
  - Setup HADOOP\_HOME to C:\hadoop-common-2.2.0-bin-master
    - Avoid nasty bug which prevents you from loading files in Spark

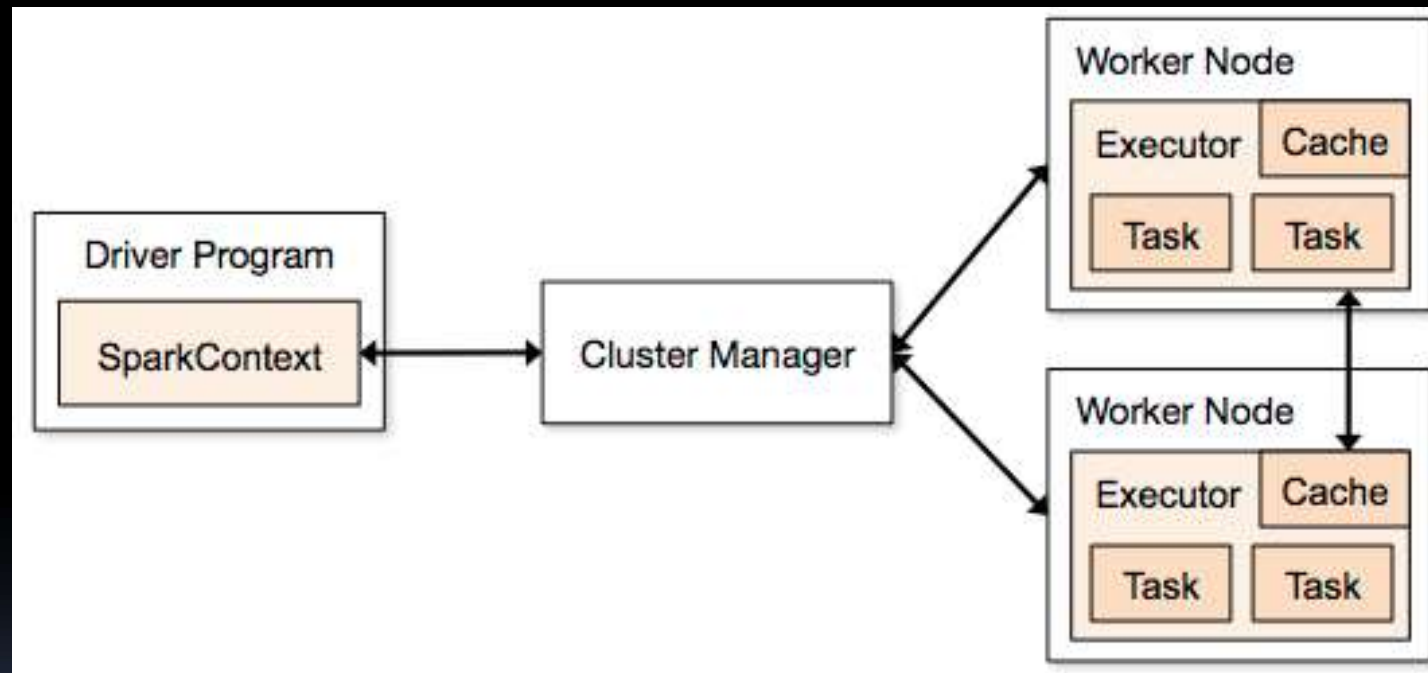
# Resources

- <http://spark.apache.org/docs/latest/programming-guide.html>
- <http://spark.apache.org/docs/latest/streaming-programming-guide.html>
- <http://spark.apache.org/docs/latest/sql-programming-guide.html>

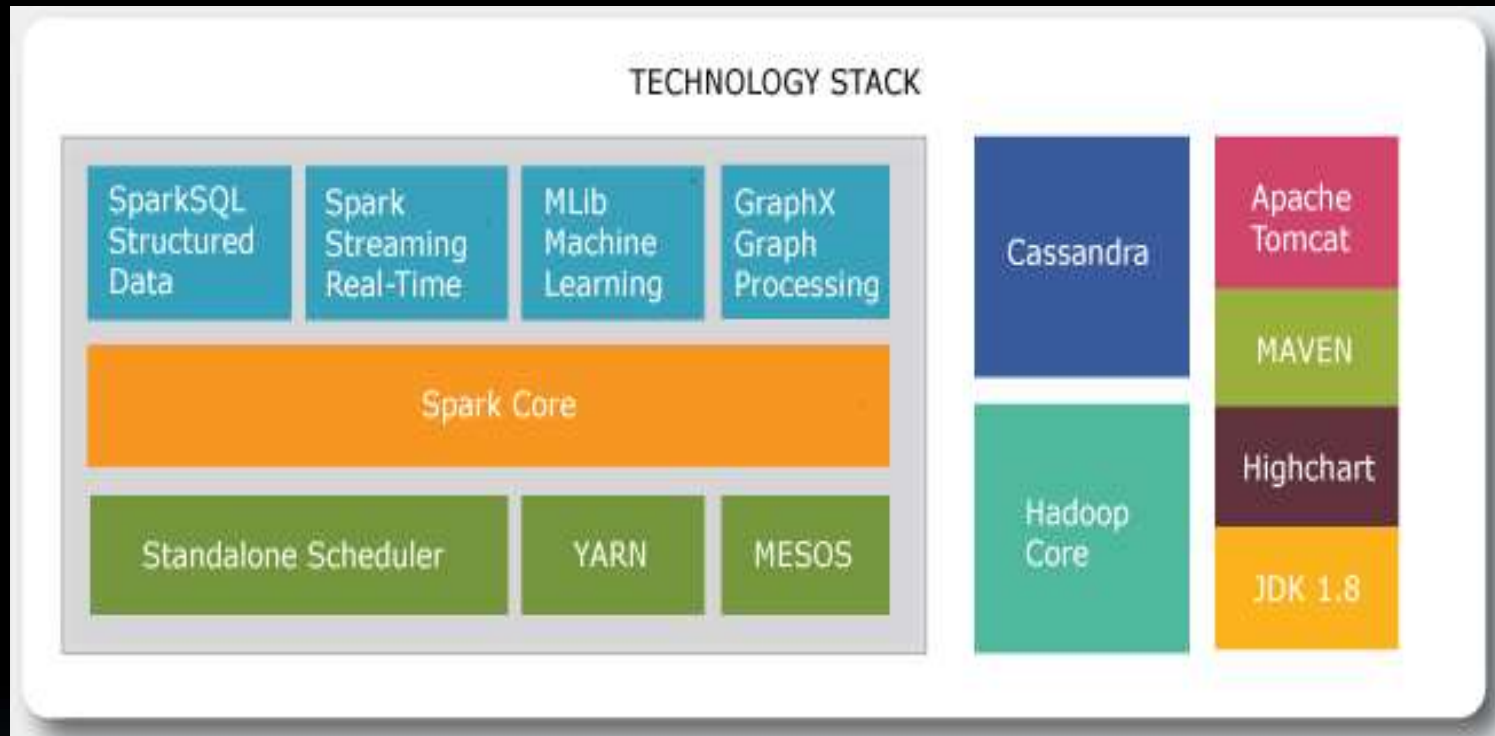
# Standalone Apache Spark Application

- Start Simple!
- Word Count = Apache Spark Equivalent of Hello World
- RDD's in practice
  - Operations
    - Transformation
      - Return a new RDD
    - Action
      - Return a result

# Spark Driver in Action



# Architectural Framework





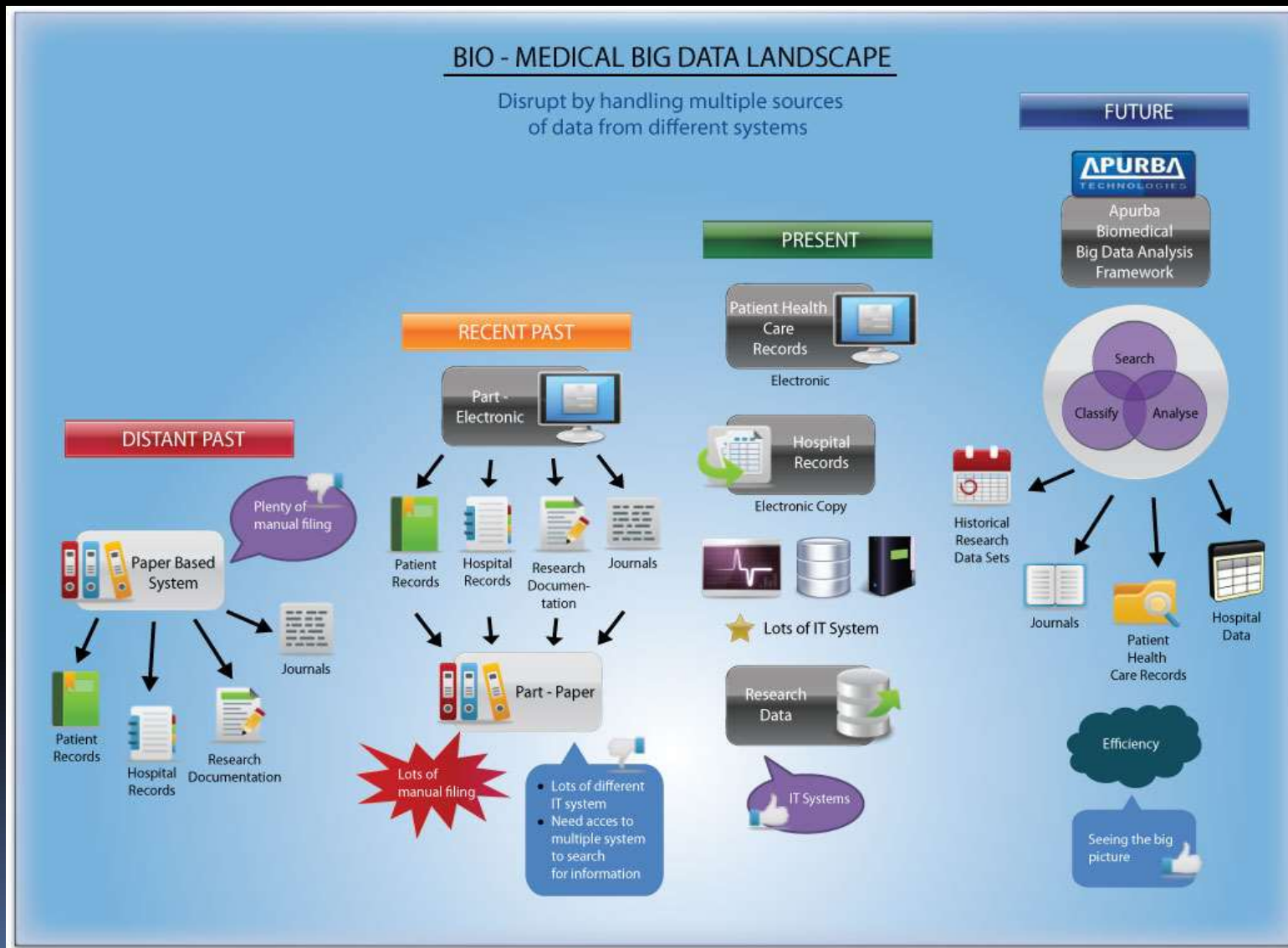
# Commonalities in demand between Medical and Financial Data Sets

- Handling of large data sets
- Structured/Unstructured data
- Data Aggregation
- Extraction of Key Performance Indicators
- Leveraging of existing code base and ecosystems

# Application in Health Care

- Assessing a patient in less than 8 minutes
- Creating a standard for evidence based medicine
- Interfacing disparate data sets
  - Structured data
    - RDMBS
  - Unstructured
    - Patient Notes
    - Doctors Notes database

# Health Care Big-data landscape



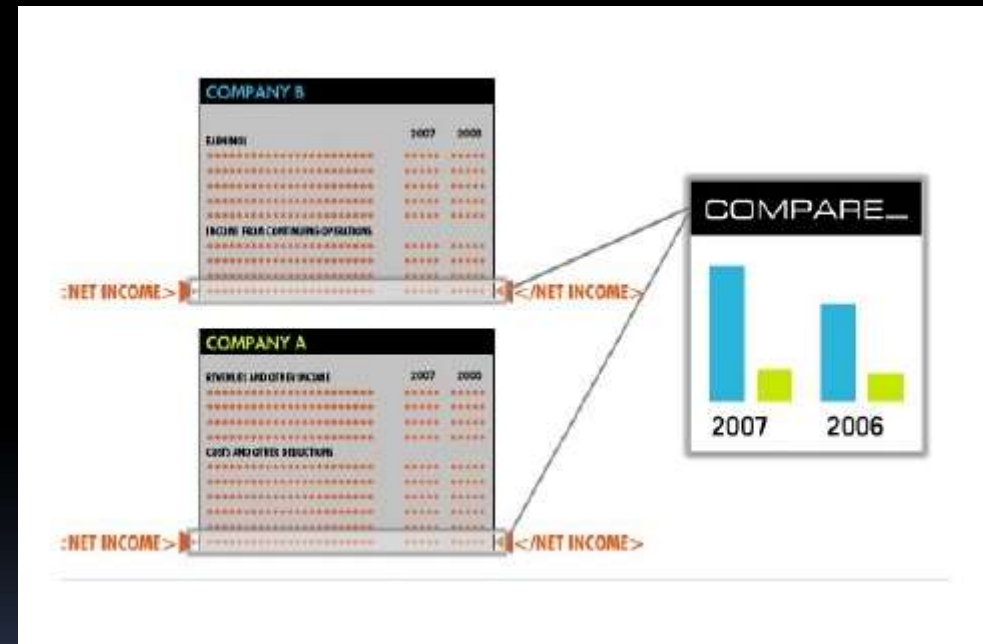


# Application in Financial Analysis

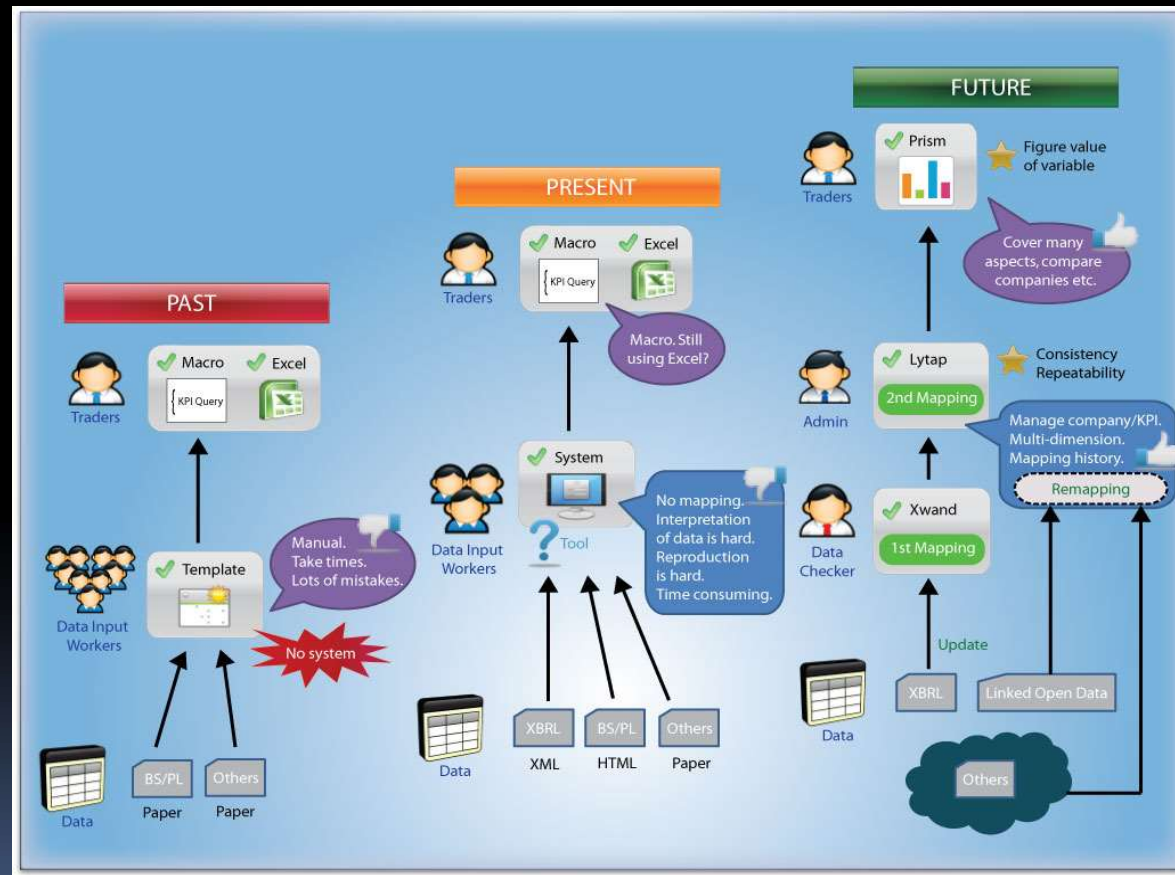
- Seizing an opportunity
  - Security Exchange Commission (SEC) repository of financial disclosures
  - Quarterly reports of every major US listed Company
- Massive of eXtensible Business Reporting Language (XBRL) format

# What does XBRL do?

- What is XBRL
  - eXtensible Business Reporting Language
  - Each line item is given data tag standardized by US GAAP and different industries
- What does it do?
  - Creates machine readable data
  - Facilitates exchange of financial data between IT Systems



# Insights from XBRL Silos



# Opportunities

- Company Valuation
- Forecasting
- Competitive Position
- Leveraging Tax Benefits
- Tax Compliance
- Detection of Tax Evasion




# Section #7: Conclusion

It's about time we wrap up!



# Conclusion

- Big-data is opening up huge possibilities in healthcare and other areas
- The market indicators are forecasting very aggressive growth of this industry
- In all predictions, this area is here to stay and will become ubiquitous in all spheres of life
- What does to us as a society is however a very open question...



Thanks for your time and  
attention!

If you want to contact me, email me at [ari@apurbatech.com](mailto:ari@apurbatech.com)