Enterprise Data Problems in Investment Banks

"BigData" History and Trend – Driven by Google

CAP Theorem for Distributed Computer System

Open Source Building Blocks: Hadoop, Solr, Storm..

3548

Hypothetical Solution using Lambda Architecture

Where "BigData" Industry is Going?

SQL + NOSQL + NEWSQL + REALTIME FOR INVESTMENT BANKS

CHARLES CAI
ASHWANI ROY

8 March 2013

Presenter: Charles Cai

- Charles Cai makes a living by designing and implementing trading and risk systems for investment banks.
- Currently a Chief Front Office Technical Architect in a global energy trading firm.

  - Twitter: @caidong
  - Linkedin: charlescai

Presenter: Ashwani Roy

- Ashwani Roy – Masters in Finance Student at London Business School and VP at a Tier 1 Investment Bank.
- Love to mix programming and Applied Mathematics to solve difficult problems in Investment Banking

  - Twitter: @Ashwani_Roy
  - Linkedin: ashwaniroy

3548

Why Finance Industry should care ?

- We care because of
  - Compliance requirements
  - Risk Management
  - Pricing
  - Rise of Machines (Ecommerce)
  - Cost Cutting
  - BTW: Twitter is also part of Market Data

Sample Interest Model / Simulations

$$dr = k(r_l - r_t)dt + \sigma * dw_t$$

A quick Monte Carlo Demo

- Demo – Computing this is functional

Some Terminology

- PV = present value = Cash flows discounted to current time
- Delta = change in price / change in interest rate
- Gamma .. Vega .. Rho .. Theta .. Vanna …. And other Greeks

Monte Carlo Simulations -Results

- <results> = func<I,j,k……
- Parallelize computation with mappers
- Save results and run reducers
- [[trade: 1 curveid: Orig PV:100 Delta:200]{ to OLAP}

  ..[trade: 1 curveid: Sim1 PV:100 Delta:200] {big data}

  ..[trade: 1 curveid: Sim2 PV:99 Delta:220]{big data}]]

Compliance

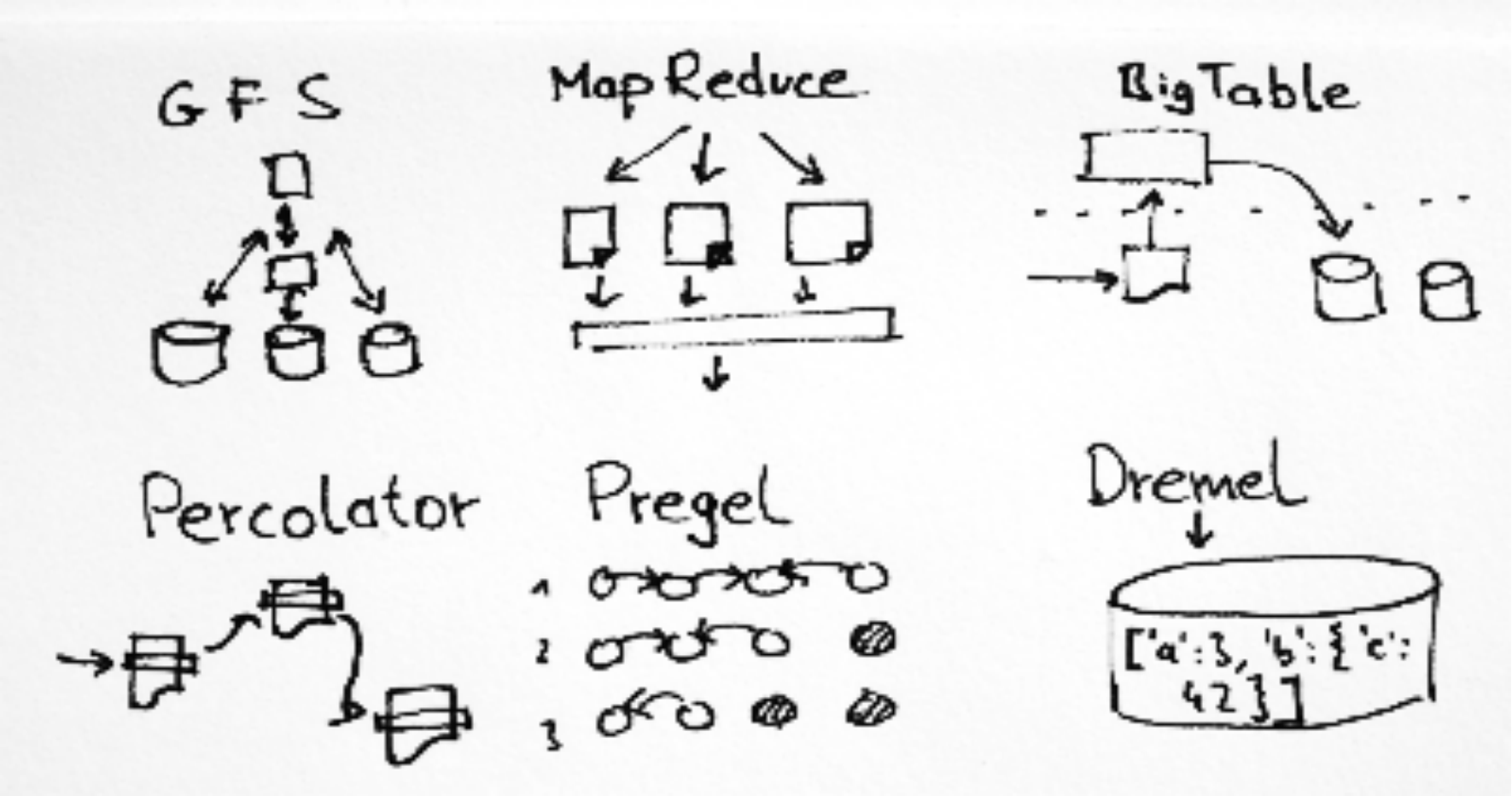
- Dodd-Frank requires >= five years records
- Fast Disaster recovery requirements (Tapes backup not acceptable)
- All Bloomberg and other chats to be saves in quick reportable form
- … Many more in Basel 3 and Dodd Frank Act

You need to

get chats for AshwaniRoy@bloomberg.net and ashwaniR@reuters.net
from the 5 years Bloomberg and Reuters log of a global investment bank of 1TB(assume 1MB/Day/Trader * 220 trading days * 1000 traders* 5 years)
for all EURUSD swaps only

….. Additional filters and aggregation requirements

Big Data Industry History: Google's Papers



Google's Big Data Papers: 2003 – 2006

GFS – Google File System

- 2003
- Distributed file system
- 3 x copies
- Commodity machines
- Colossus (2012)

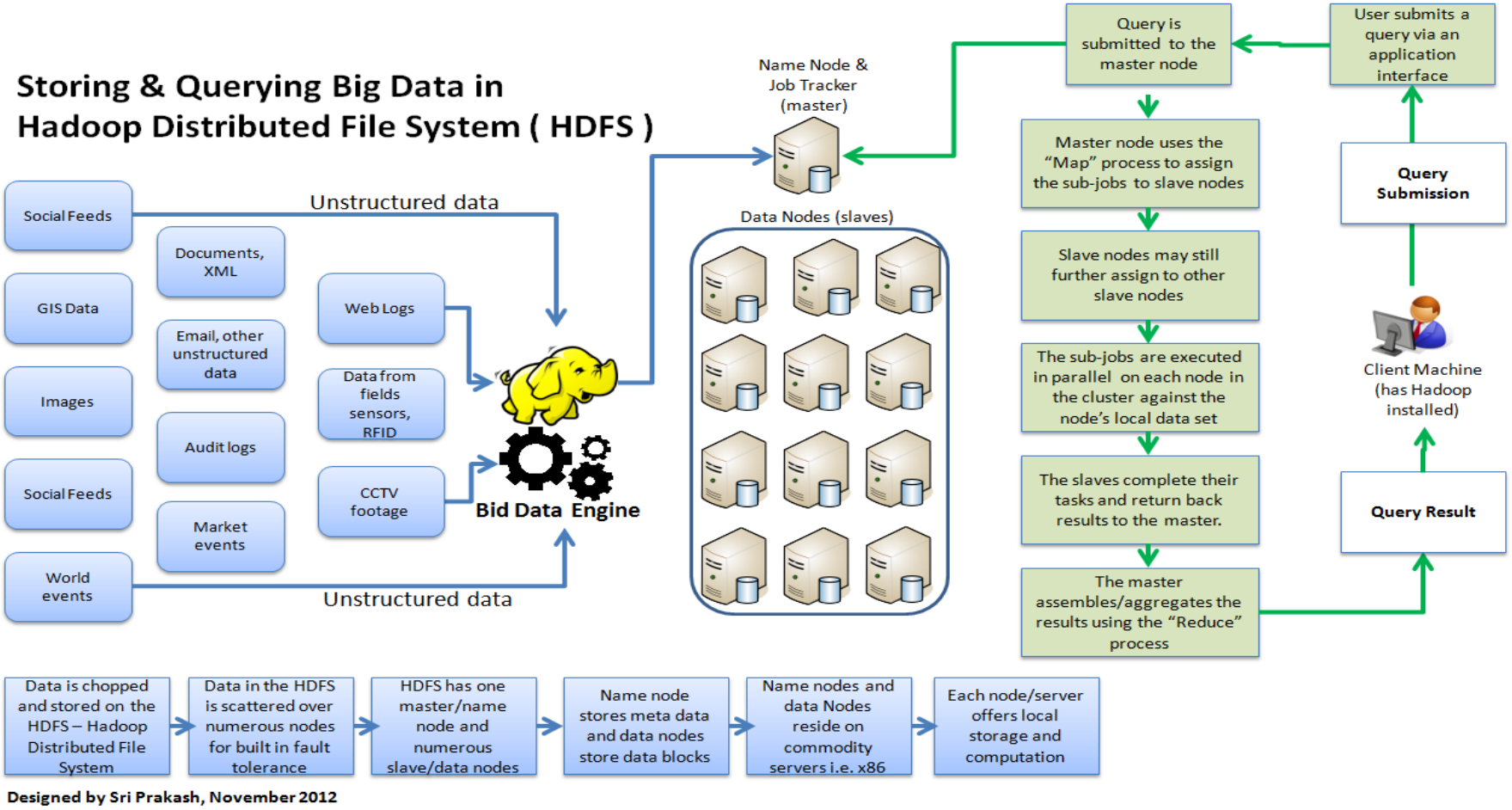
MapReduce

- 2004
- Input → Map → Partition → Compare → Shuffle → Sort → Reduce → Output

BigTable

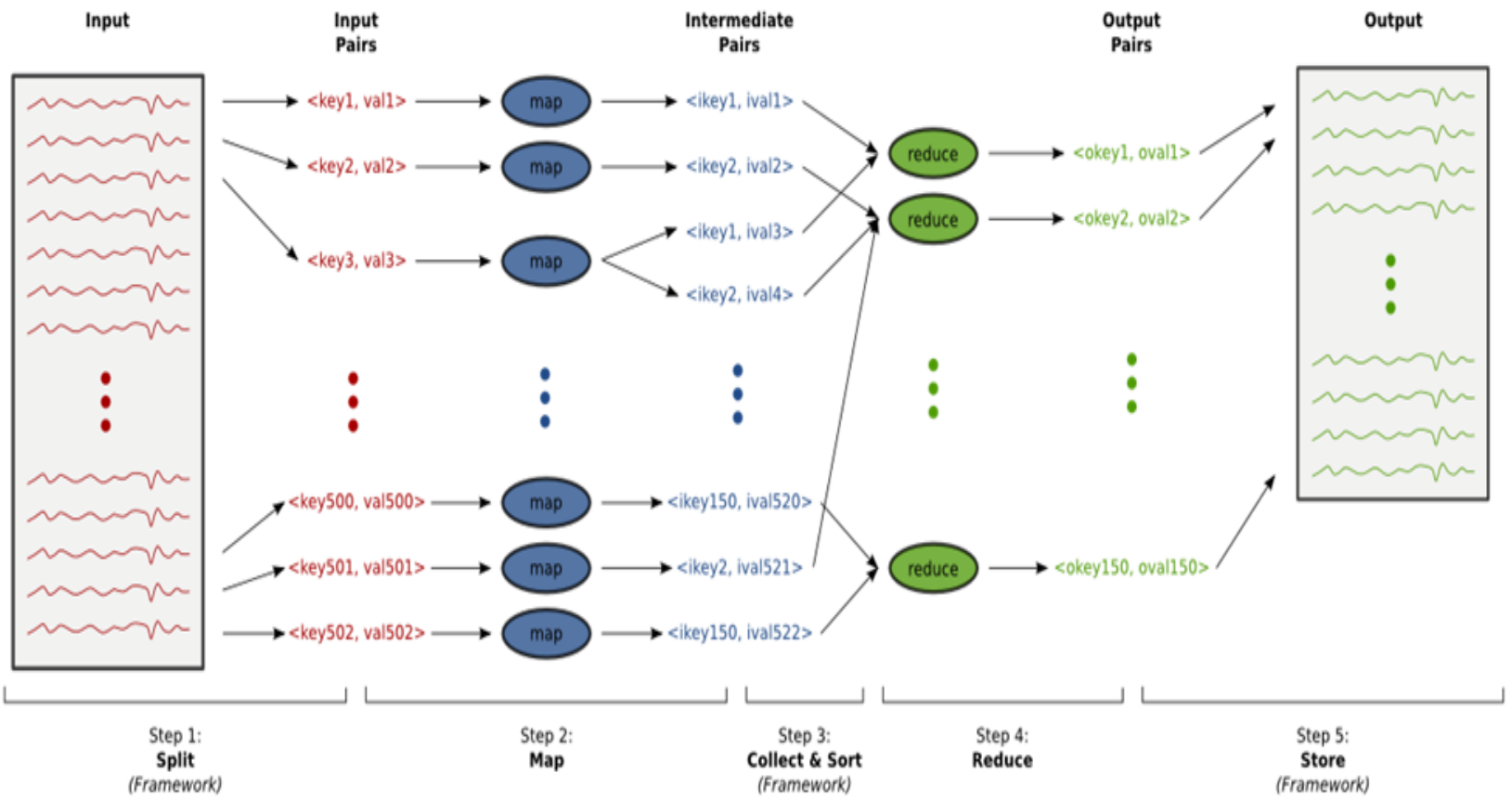
- 2006
- Distributed Key-Value column-family based database

Hadoop Distributed File System (HDFS)

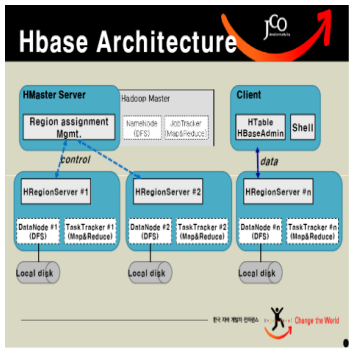


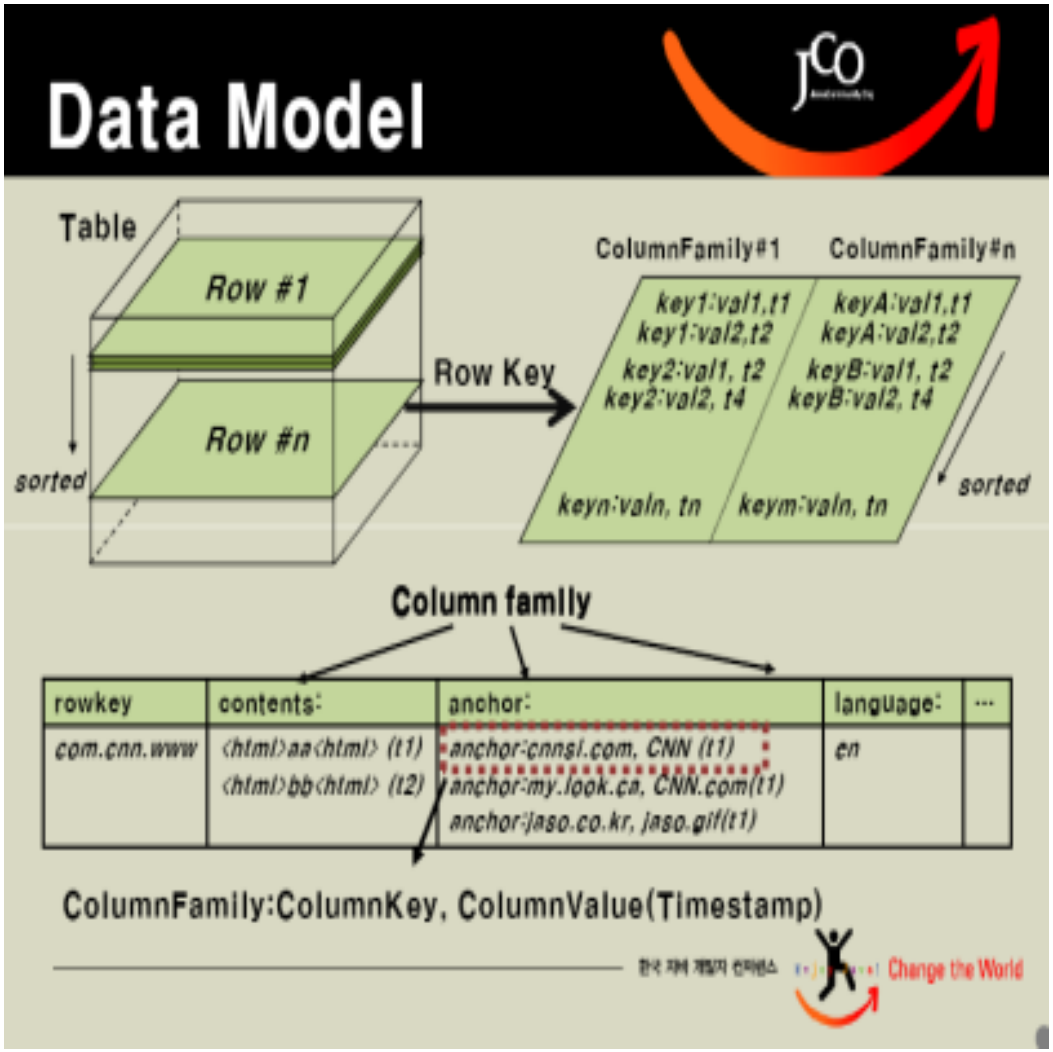
- http://ecomcanada.files.wordpress.com/2012/11/hadoop-architecture.png

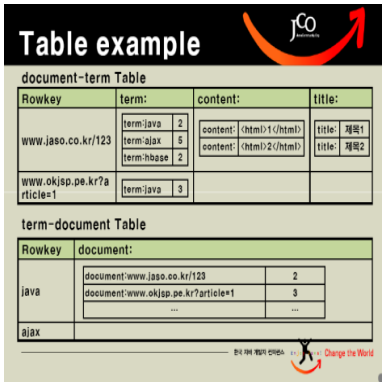
Google's MapReduce Programming Model



Apache Hbase: Column Family Distributed K-V Store







Google's Big Data Papers 2: 2010 - now

Percolator

- 2010
- Incremental update/ compute
- built on BigTable
- Adds transactions, locks, notifications
- SPFs: "Stream Processing Frameworks" + underlying database

Dremel

- 2010
- Online analytics and visualization
- SQL like language for structured data
- Each row is JSON object – in protobuf format
- Column based
- **Spanner (2012), BigQuery, F1**

Pregel

- 2010
- Scalable graph computing
- Worker threads → nodes → parallel "superstep" → messages → nodes → Aggregator/Combiners (global statistics)
- **PageRank,** shortest path, bipartite matching

YAHOO! S4 distributed stream computing platform — Apache incubator

Storm — Distributed and fault-tolerant realtime computation

Apache Drill — Apache incubator

cloudera — Ask Bigger Questions — **Impala**

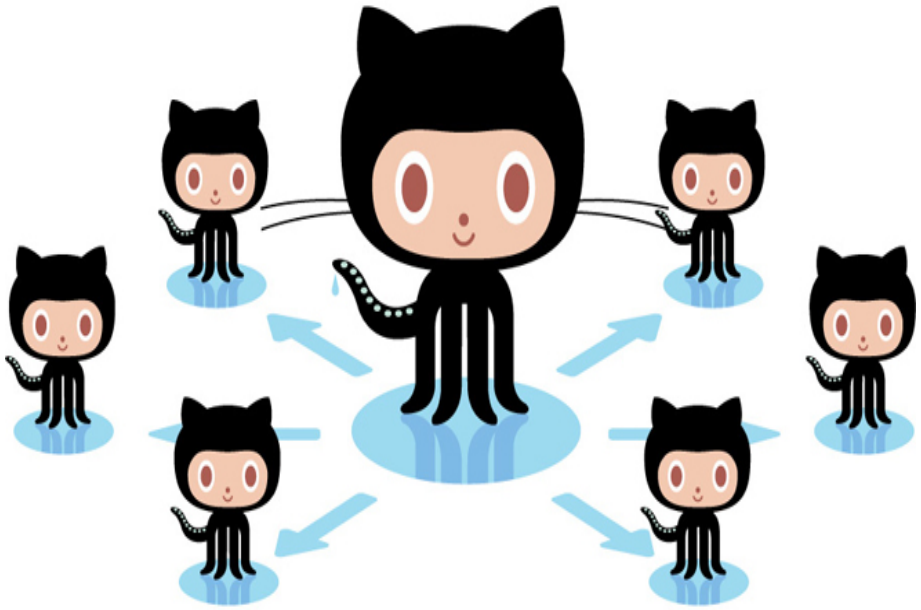
Hortonworks — **Tez/Stinger**

Apache Giraph — Apache incubator

TITAN

Microsoft Trinity

Unstructured Data: Index/Search Engine



- Github Code Search: 17 TB

Apache Lucene/SOLR

- Open Source Indexing and Search Engine
- 4,000+ Enterprise users
  - IBM, HP, Cisco
  - Apple, Linkedin
  - Wikipedia
  - CNet, Sky
  - Twitter

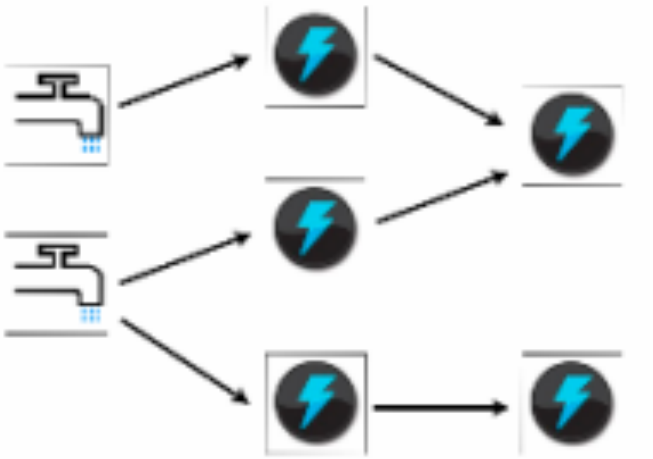
What's Next for Hadoop? Real-time!

Storm
Distributed and fault-tolerant realtime computation

Background

Nathan Marz

Backtype, Storm, and Twitter



http://storm-project.net/

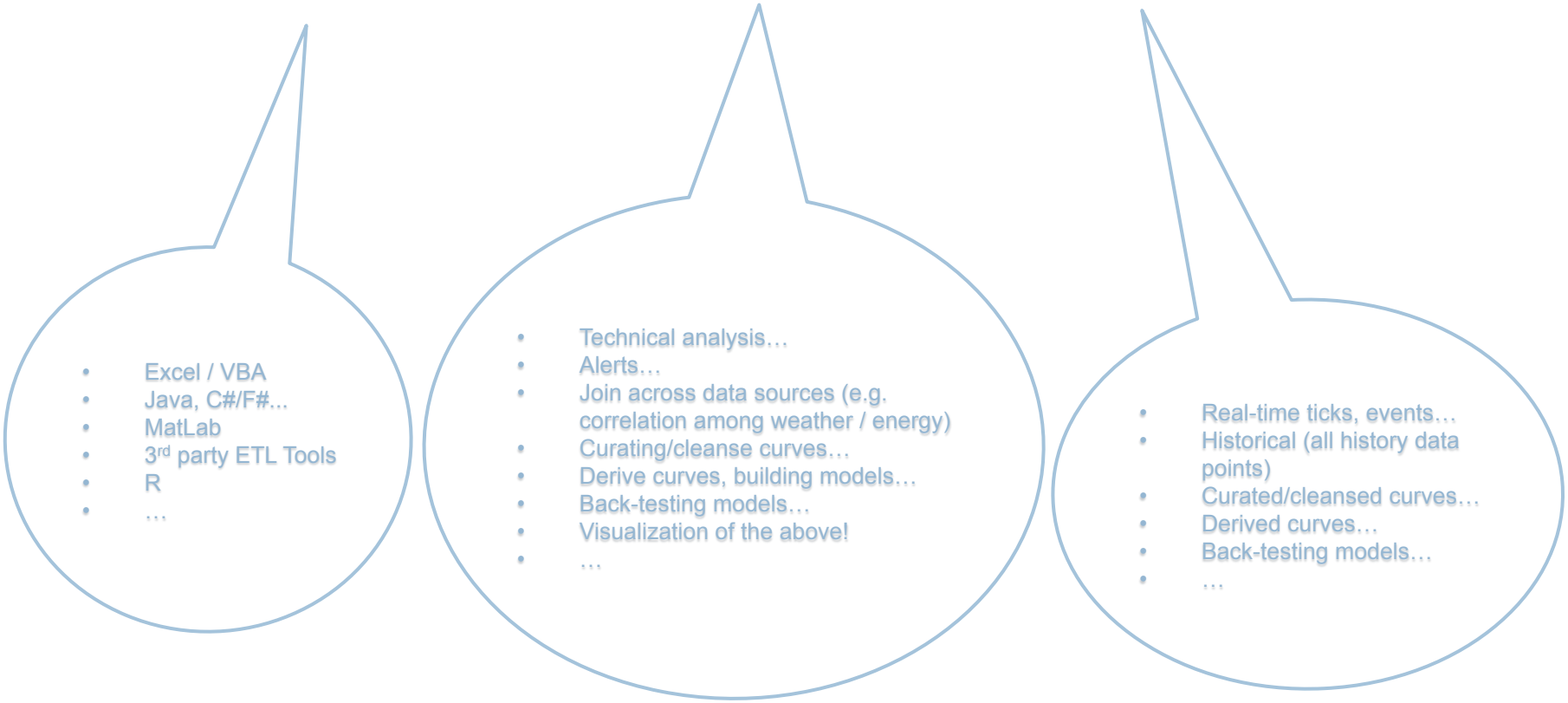
http://www.manning.com/marz/

Some more use cases

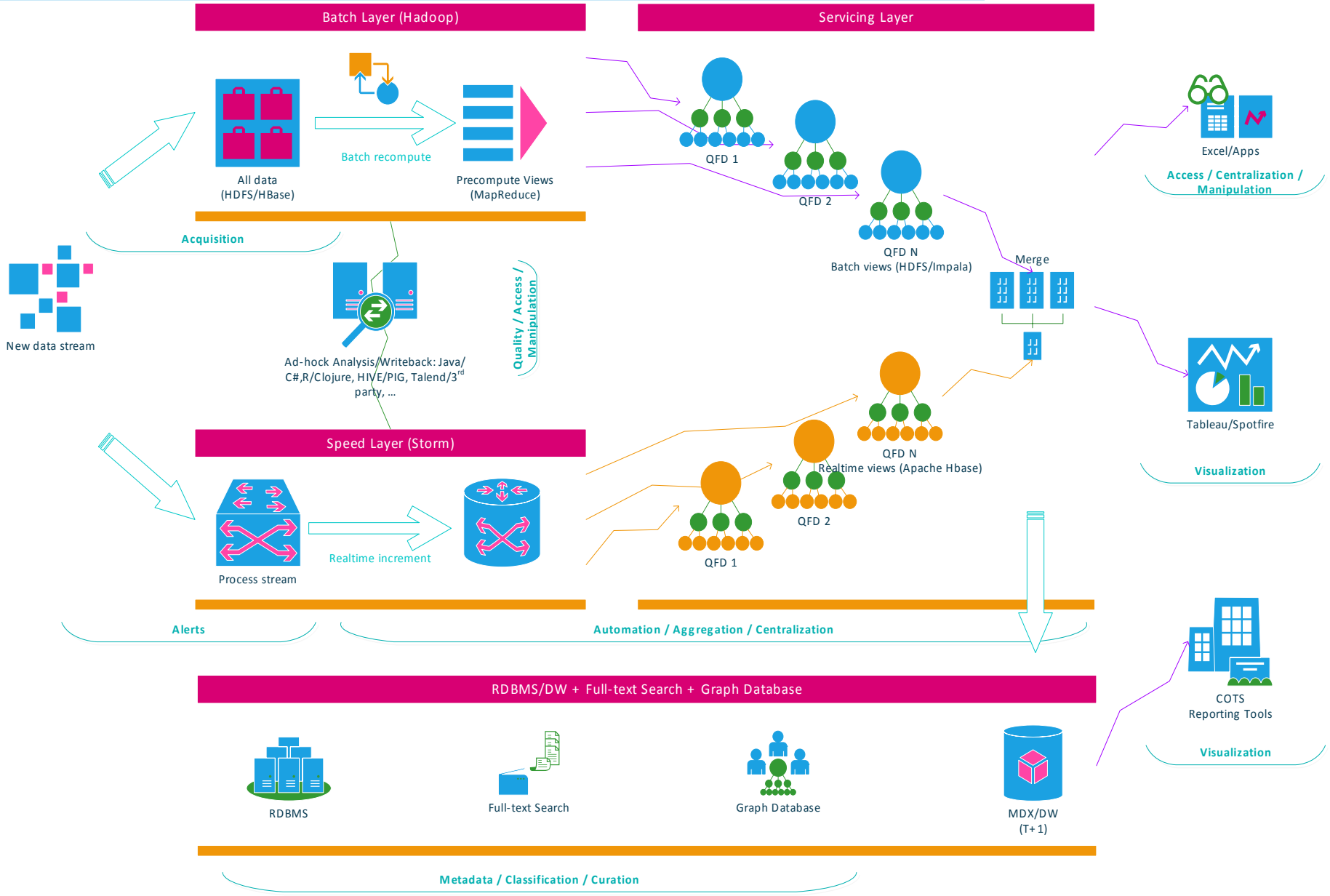
- Save money to save your jobs
- Save money to your firm can do more
- E Commerce is norm…
- Market sentiment analysis cannot be relied on using "Bloomberg's sentiment analysis" only
- .. Add some more

"Lambda Architecture" – Nathan Marz, BackType/Twitter

- query = func (data, …)



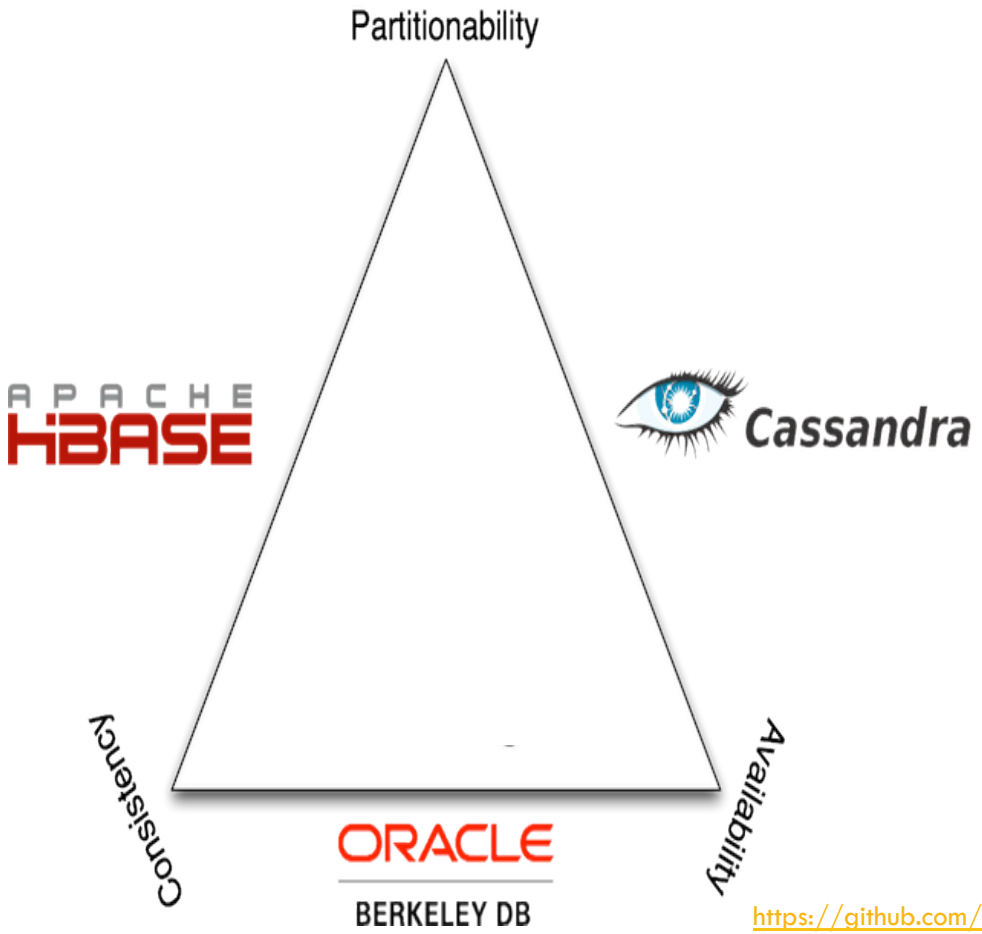
Lambda Architecture : query = func (data, ...)



Online resources and alternative stacks

- An Introduction to Data Science.PDF – Free e-book on Data Science with R under Creative Commons Licenses
- Berkeley Data Analytics Stack (Open Source: Mesos – cluster management, Spark/Streaming – cluster computing, Shark-SQL/DW)
- Learning Statistics with R, Free Big Data Education: Advanced Data Science
- DataStax Enterprise (Apache C*/Cassandra, Apache Hadoop, Apache Solr…)
- An example "lambda architecture" for real-time analysis of hashtags using Trident, Hadoop and Splout SQL
- Nathan Marz (BackType, acquired by Twitter) Big Data Lambda Architecture
- Open source clustered Lucene: elasticsearch used by GitHub (17 TB code)

Distributed Computing System: CAP Theorem



Consistency

- all nodes see the same data at the same time

Availability

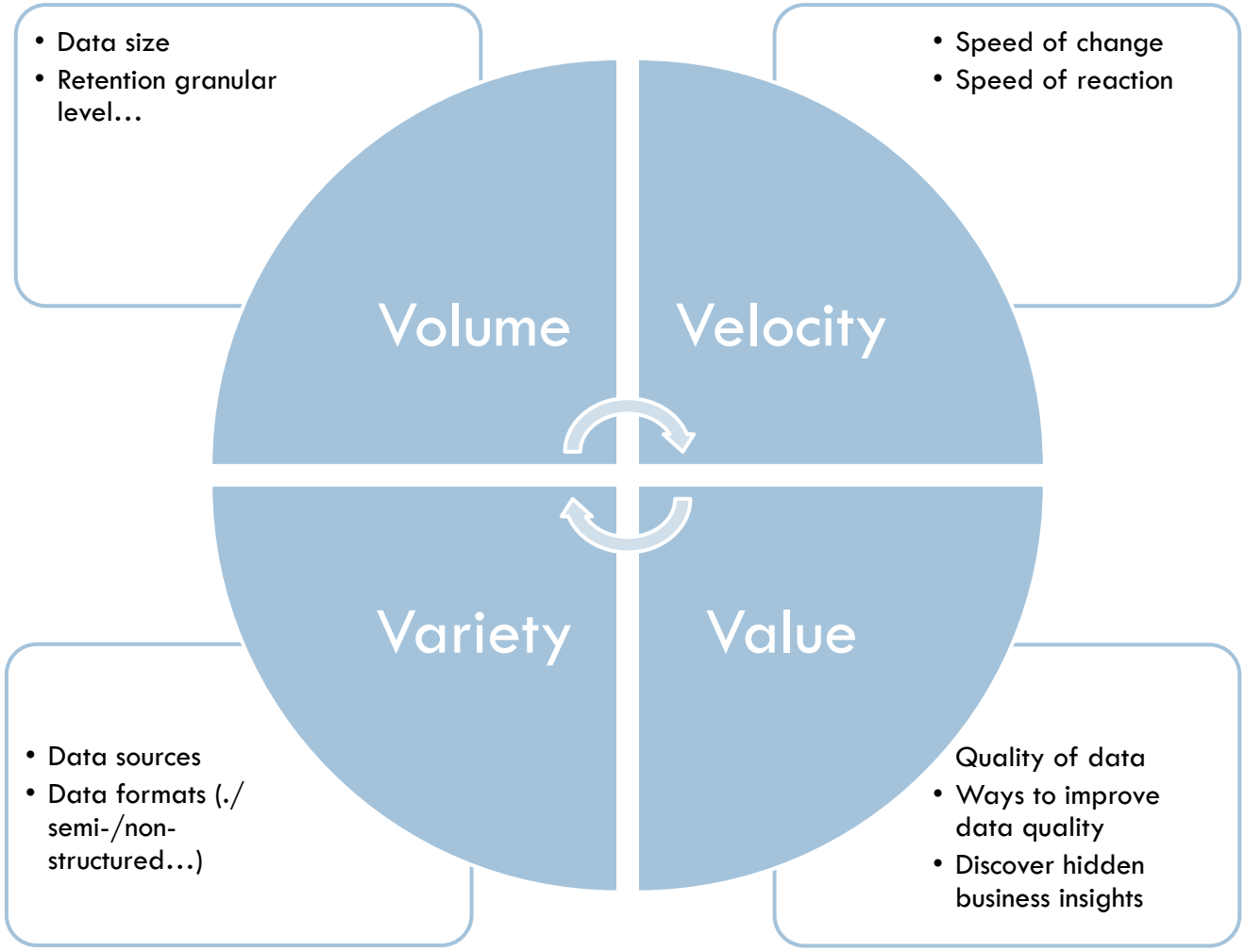
- a guarantee that every request receives a response about whether it was successful or failed

Partition tolerance

- the system continues to operate despite arbitrary message loss or failure of part of the system

https://github.com/thinkaurelius/titan/wiki/Storage-Backend-Overview
http://en.wikipedia.org/wiki/CAP_theorem
http://www.infoq.com/articles/cap-twelve-years-later-how-the-rules-have-changed

“Lambda Architecture”: Enterprise Data



“Lambda Architecture” – Nathan Marz, BackType/Twitter

- Design Principle:
  - Human fault-tolerance
  - Immutability
  - Pre-computation
- Lambda Architecture:
  - Batch Layer
  - Serving Layer
  - Speed Layer
- Technology Stack
  - Apache Hadoop/HBase/Clou
  - Twitter Storm