# Big Data Overview

## Presented to DAMA

Lynn Hedegard
Lynn.Hedegard@us.ibm.com
Technical Sales Specialist
West Region

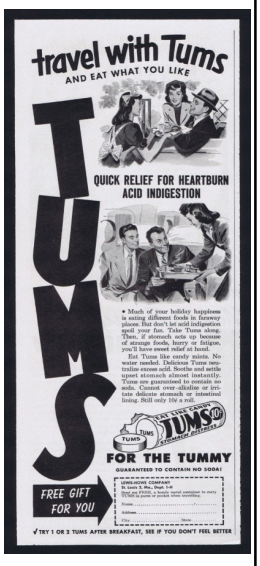
28th of May, 2015



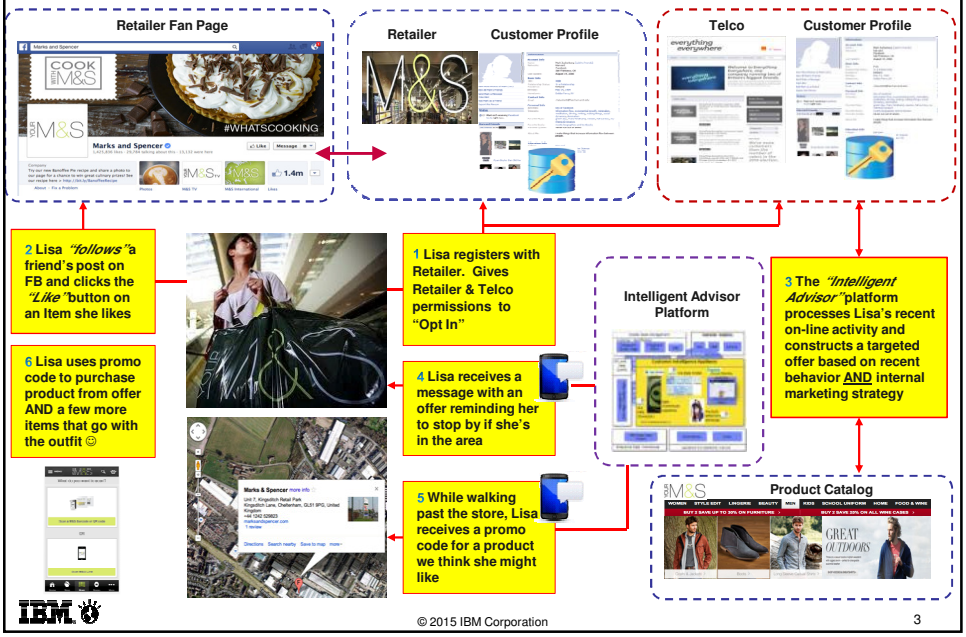
## When Was CRM Invented?

- As early as the 1930's some marketing companies were using coupons to track and measure response to printed advertising
- The coupon was usually located at the bottom of a full page ad.
- They said they were going to give you something for free…

  … but they were actually measuring the effectiveness of the ad.
- Some of these ads are now collector's items



## Real-Time CRM in the Social World (Meet Lisa)







## How Did They Do That?

How Did
They Do That?

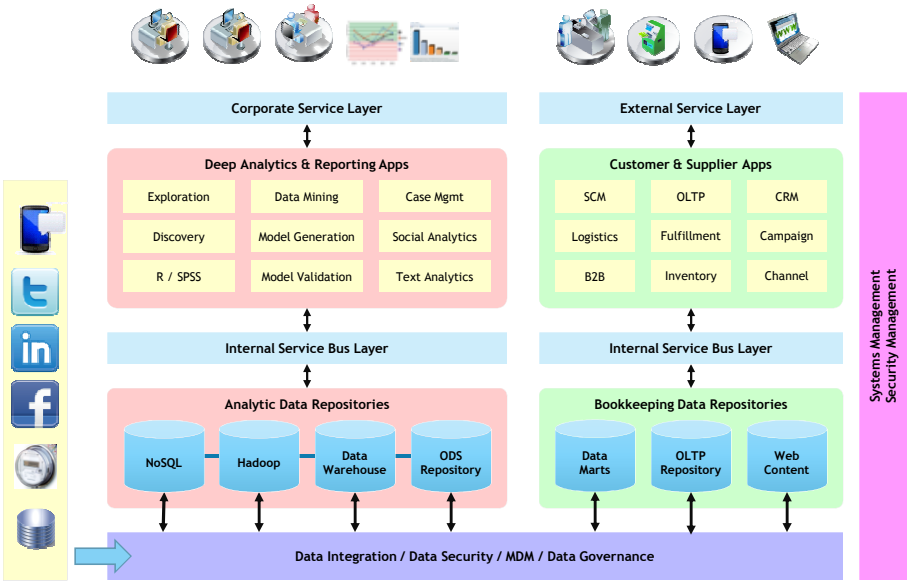
## Technologies Used in Previous Example

- **They used Legacy Technology**
  - Online Transaction Processing Systems (OLTP)
  - Operational Data Stores (ODS)
  - Data Warehouse (DW)
- **They used "Newer" Technology**
  - Telecommunication
  - Mobile Computing
  - Wireless
  - Business Intelligence
- **They used Big Data Technologies**
  - Hadoop
  - NoSQL Database Systems
  - Data Science
  - Real-Time Analytic Processing

We will focus on this topic of Big Data




## Reference Architecture for the On-Line Enterprise



## Problem Statement

# Problem Statement

(or why do we need Big Data)




## Problem Statement — Complex Business Environment

- The Local Environment is Complex:
  - A single large retail store (1.5 million SKUs)
  - Large manufacturing floor (~6 million parts)
  - Vegas Casino (20 million card carrying customers)
- The Global Environment is Complex:
  - The number of variables affecting business performance is huge.
  - US citizens (*source: google population*)
    - 300+ Million total
    - (21M+ teenagers) + (40M+ in their 20's) (that's a lot of calls & text messages!)
  - The interrelationships between these variables is very complex (e.g., $N^2$ problem)
    - Multiple customer touch points
    - Multiple suppliers & distribution methods
    - Market forces (cost of raw goods & services, pricing dynamics, supply/demand)
- Working Premise: Few people in the enterprise can make "good" Operational Decisions — consistently & quickly
  - Few people can *"see"* all the necessary data.
  - Few people can *"analyze"* all the necessary data.
  - Few people understand all the inter-relationships between business variables.

Businesses can no longer tolerate inconsistent Business Processes

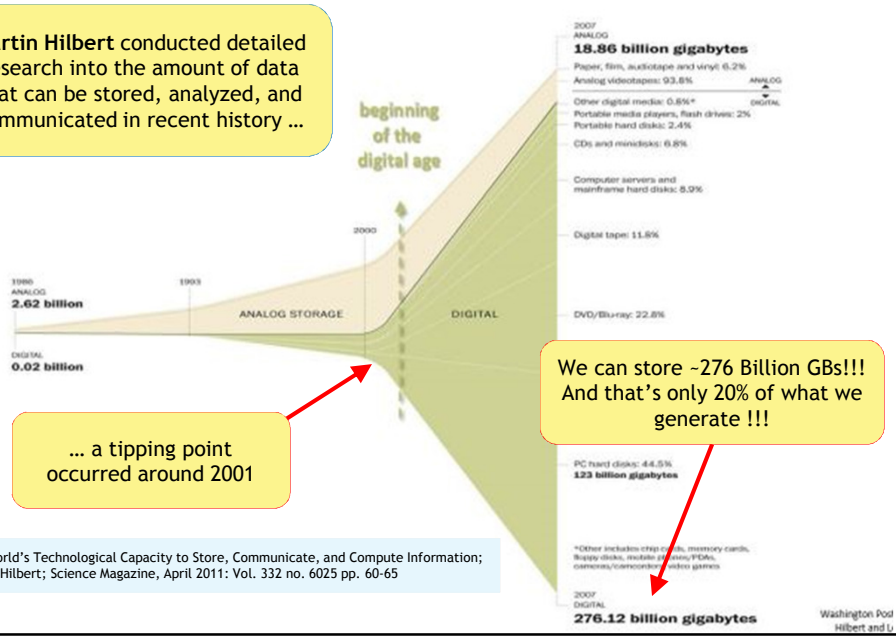
But wait …

… We're just getting started




## Global Capacity to Store & Communicate Information

**Martin Hilbert** conducted detailed research into the amount of data that can be stored, analyzed, and communicated in recent history …




… a tipping point occurred around 2001

We can store ~276 Billion GBs!!! And that's only 20% of what we generate !!!

The World's Technological Capacity to Store, Communicate, and Compute Information; Martin Hilbert; Science Magazine, April 2011: Vol. 332 no. 6025 pp. 60-65

## How Did the Big Data Wave Get Started?

**Prior to 2000**, not many people were using the internet.

- Modems were still slow.
- Video resolution was poor.
- Computer interfaces were clumsy.
- The average person might use e-mail, but not much more

- **Xerox** created the *"window interface"* and the *"mouse"* at the **Palo Alto Research Center (PARC Place)**, but not many people even know about it.

- I had a large black & white X-Terminal on my desk
- I was in heaven!!!



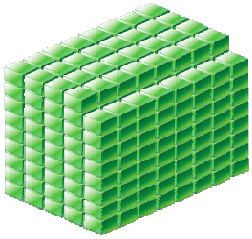

## Society Experiences a "Tipping Point" in 2001

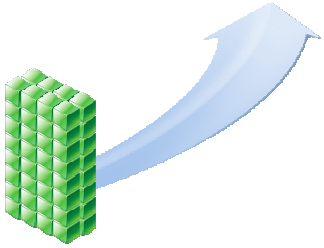
- Significant advances in technology cause a tipping point
  - Telecommunication capabilities advanced to the point where an *"average person"* could tolerate using modems.
  - We move from command-line interfaces to window/mouse based interface (created by Xerox).
  - The average person starts using the internet for …
    - E-Mail, Instant Messaging, Shopping,
    - and of course … Web Surfing.
  - Most adults (and many teenagers) have a cell-phone
  - Internet companies like, AOL, Yahoo, MicroSoft, & Google get overwhelmed with large quantities of DATA
  - **Conventional database system were ill-suited of storing this data**
    - Data was semi-structured
    - Data was not clean
    - And there was a LOT of this data.

## Growth of New Data

- 200 million emails per minute
- 2.5 million items per minute
- 2 million searches per minute
- 300,000 tweets per minute
- 220,000 photos per minute
- 5 TBs per flight
- > 1 PB per day from Gas Turbines

**2020**: *40 Zetta-Bytes*



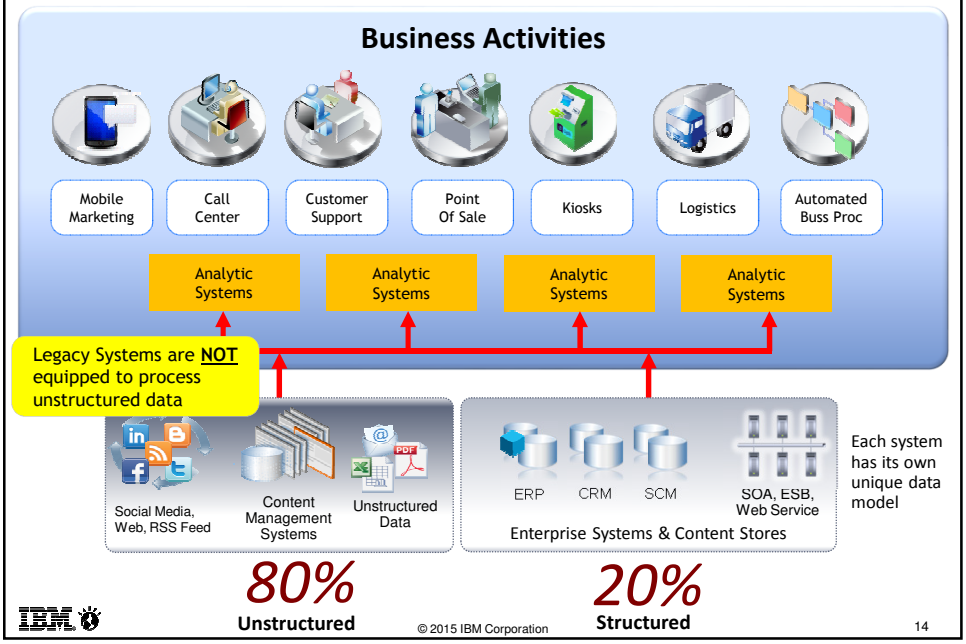


**2012**:
- 2.5 Exa-Bytes / Day
- 2.8 Zetta-Bytes Total





## 80% of Data Growth is Unstructured Data

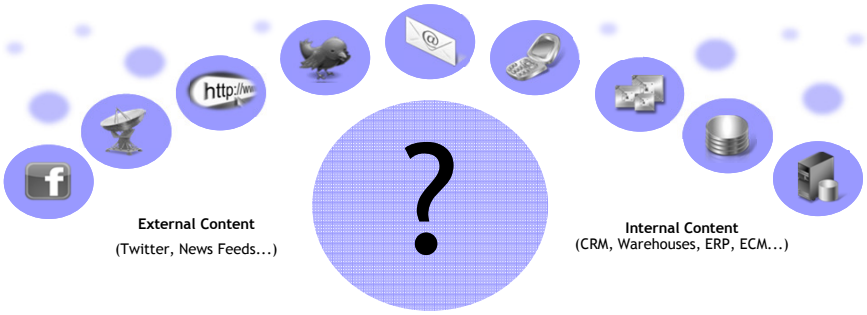



*80%* **Unstructured**

*20%* **Structured**

## Data is Everywhere

**Data is everywhere, but under-utilized at the point of impact**




**External Content**
(Twitter, News Feeds...)

**Internal Content**
(CRM, Warehouses, ERP, ECM...)

"I can't find the right answers fast enough to support my customers."

"I need to know what complications to expect, if I use this new medicine before the surgery."

"I know there's value in my data. How do I convert all this information into economic value?"


"I don't know what I don't know!!!

Where is my business exposed?"

"It was a cold day in the city. The rain fell like tears on the pavement. I had a bad feeling about this."



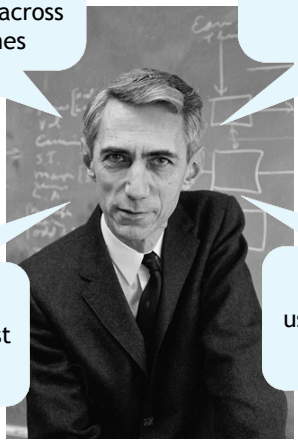

## Big Data Discussion

Big Data Discussion

## Lets Make a Business Case for Big Data

Back in 1948 … I showed you how to digitize voice & music (e.g., MP3, Compact Disk) **AND** how to send all this stuff across telecommunication lines

… I also showed you the **BUSINESS VALUE** of Data & Analytics



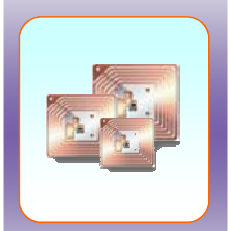
Next, I collaborated with Ed Thorp to create the first wearable computer …

… We went to Vegas and used the wearable computer to win at Blackjack

Claude Shannon (Father of Information Theory)
**1948: First Paper on Information Theory**




## We Need To Think Beyond Traditional Sources Of Data

| Transactional & Application Data | Machine Data | Social Data | Enterprise Content |
|---|---|---|---|
|  |  |  |  |
| Volume<br>Structured<br>Throughput | Velocity<br>Semi-structured<br>Ingestion | Variety<br>Highly unstructured<br>Veracity | Variety<br>Highly unstructured<br>Volume |

## We Need To Think About the Attributes of Data

| Volume | Variety | Velocity | Veracity |
|---|---|---|---|
|  |  |  |  |
| **Data at Scale**<br>Terabytes,<br>to Petabytes<br>to Exabytes,<br>and Beyond | **Many Data Types**<br>Structured,<br>Unstructured,<br>Sensor Data,<br>Text, Multimedia | **Data in Motion**<br>Streaming data<br>Operational Decisions<br>Sub-Sec Reaction time | **Data Uncertainty**<br>Data Governance<br>Name Management<br>Data Lineage<br>Annotation |

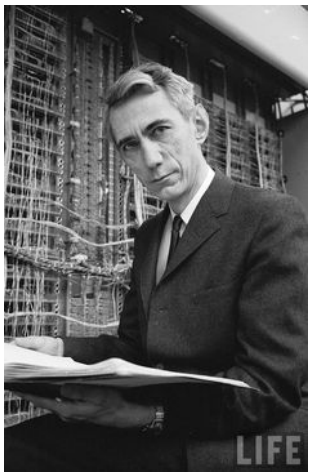
Velocity IS the game changer:
It's NOT just how fast data is produced or changed,
The speed at which can be received, understood, and processed is key




## Merging the Traditional and Big Data Approaches

| Traditional Approach<br>*Structured & Repeatable Analysis* | Big Data Approach<br>*Iterative & Exploratory Analysis* |
|---|---|
| **Business Users**<br>Determine what question to ask | **IT Group**<br>Delivers a platform to enable creative discovery |
| **IT Group**<br>Structures the data to answer that question | **Business Users & Data Scientists**<br>Explore what questions could be asked |
| Monthly sales reports<br>Profitability analysis<br>Customer surveys | Brand sentiment<br>Product strategy<br>Maximum asset utilization |

## Information & Uncertainty




*"Information is the resolution of uncertainty"*
Claude Shannon (Father of Information Theory)
**1948: First Paper on Information Theory**

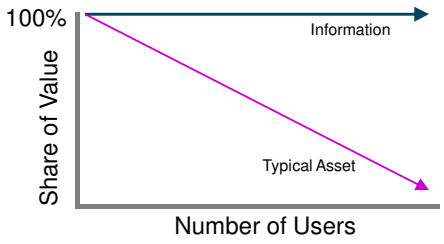
| | |
|---|---|
| 1 in 3 | Business Leaders frequently make decisions based on information they don't trust, or don't have |
| 1 in 2 | Business Leaders say they don't have access to the information they need to do their jobs |
| 83% | of CIOs cited "Business intelligence and analytics" as part of their visionary plans to enhance competitiveness |
| 60% | of CEOs need to do a better job capturing and understanding information rapidly in order to make swift business decisions |

**Analysis of Data, and Communication of insights reduces uncertainty.**

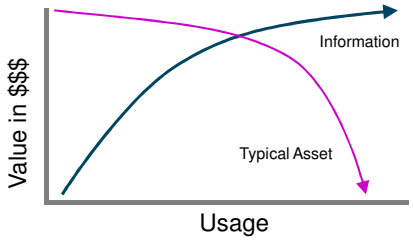



## Measuring the Value of Data — (The 7 Laws)*

- **Law One**: Information is Infinitely Shareable.



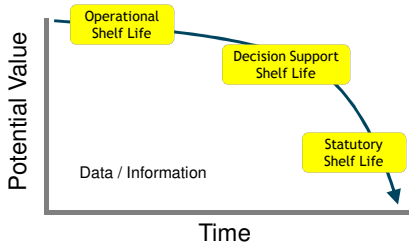

- **Law Two**: The value of data increases with its use.
- Most other assets *"wear out"* with extended use.



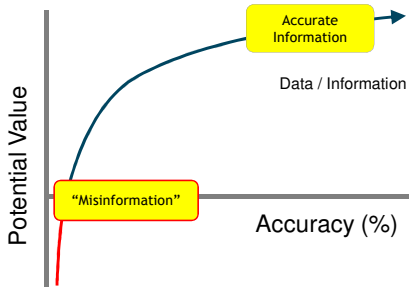

* Measuring The Value Of Information: An Asset Valuation Approach; Moody, Daniel & Walsh, Peter European Conference on Information Systems (ECIS'99)

## Measuring the Value of Data (cont.)

- **Law Three**: The value of data tends to depreciate over time.
- *"Old News"*




- **Law Four**: The value of data Increases with Accuracy.
- In fact, **inaccurate** data has a **NEGATIVE** value.

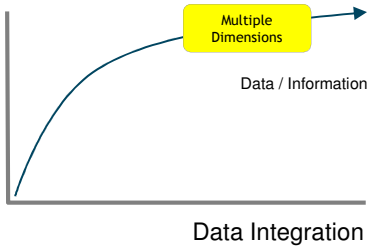




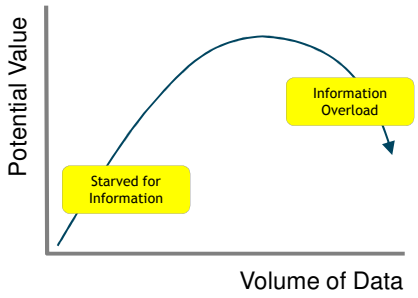


## Measuring the Value of Data (cont.)

- **Law Five**: The value of data increases when combined with other data.
- Combined value may exceed the sum of the parts



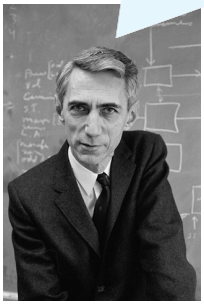

- **Law Six**: More data is **NOT** always a good thing
- *"Information Overload"*




- **Law Seven**: Information does not Deplete when used.

## Data & Information is a Valuable Asset

$$Growth = E_e[log^d W] - H(E|\vec{G}) - D_{KL}(\vec{P}(e|g)\|P(e|m)) - I(E;\vec{G})$$



Huh ?!?!?

- Growth of the Enterprise is proportional to the amount of DATA **acquired** & **analyzed**, and how well you **communicate** that insight throughout the Enterpris
  (i.e., A reduction in Entropy)
- *"The more you know … the more you grow"*
- Just look at the market value of companies that capture extreme volumes of data. (QED)
- **Key Point**: Data, and the Insights gleaned from analysis of that data is one of the most valuable assets in you Enterprise
- If you're **NOT** gathering data about your business processes and other related events, you're throwing money away!!!




## A Profile for Big Data

- Information Flow
  - **Human generated data**
    - Internet Search (Google, Yahoo, Bing)
    - Social Media Posts, Tweets, Blogs, etc.
  - **Machine Generated Data**
    - Machine Generated Logs (Linux, Network, Manufacturing, etc.)
    - Internet of Things (IOT) (Telematics, Appliances, etc.)
    - Infrastructure (e.g. Wireless, Toll Roads, Cameras, etc.)
- Information Retention
  - Amount of data retained doubles every 3 years
  - 2.5 Exabytes in 1986 (~1% digitized)
  - 300 Exabytes in 2007 (~94% digitized)
  - **About 80% of data that has been captured will be discarded !!!**
- Information Analysis
  - We can only analyze a small portion of the data retained
  - The majority of data is Semi-structured –or- non-structured

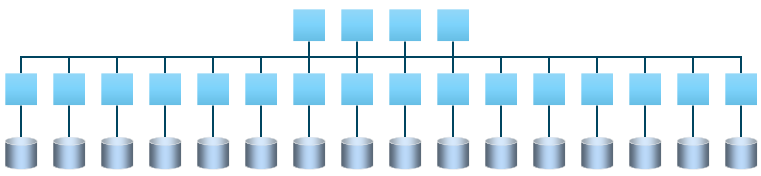
## What is Hadoop

What Is Hadoop




## Two Key Elements of Hadoop

- **Parallel Execution Framework** (MapReduce)
  - How Hadoop understands and assigns work to the nodes (machines)
- **Hadoop Distributed File System (HDFS)**
  - Where Hadoop stores data
  - The file system spans all nodes in a Hadoop cluster
  - Links together the file systems on many local nodes to make them into one big file system



## Understanding MPP

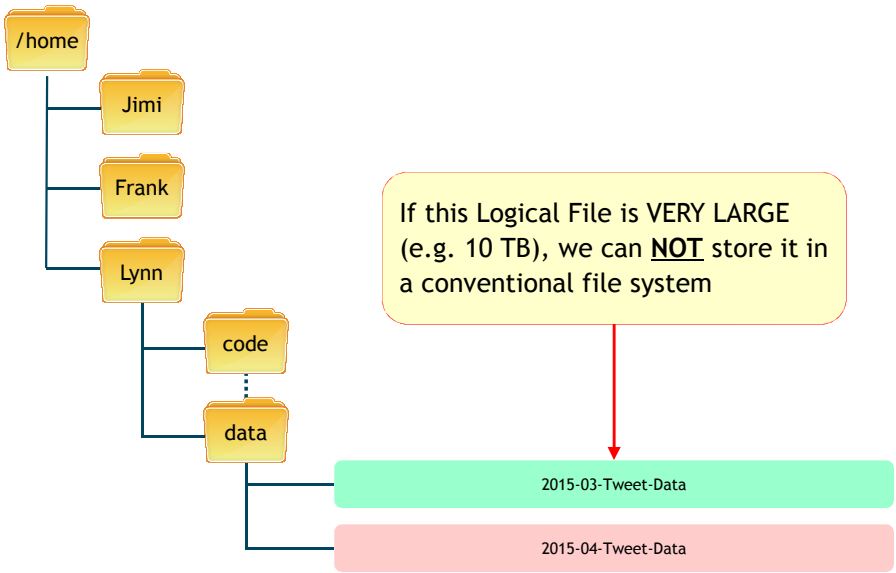
- Imagine I give you the phone book for Los Angeles
- I ask you to make a list of all the entries where…
  - First Name = “John”
  - Area Code = “323”
- You will need to look at ALL of the pages sequentially

- Now imagine we have 10 clones ready to do the same job
- We divide the phone book into 10 equal size sub-sections
- The “team” operates in parallel.
- The “team” will be able to make the list 10 times faster
- Note:
  - Data Warehousing, Hadoop, & NoSQL benefit from MPP
  - OLTP does NOT benefit from MPP

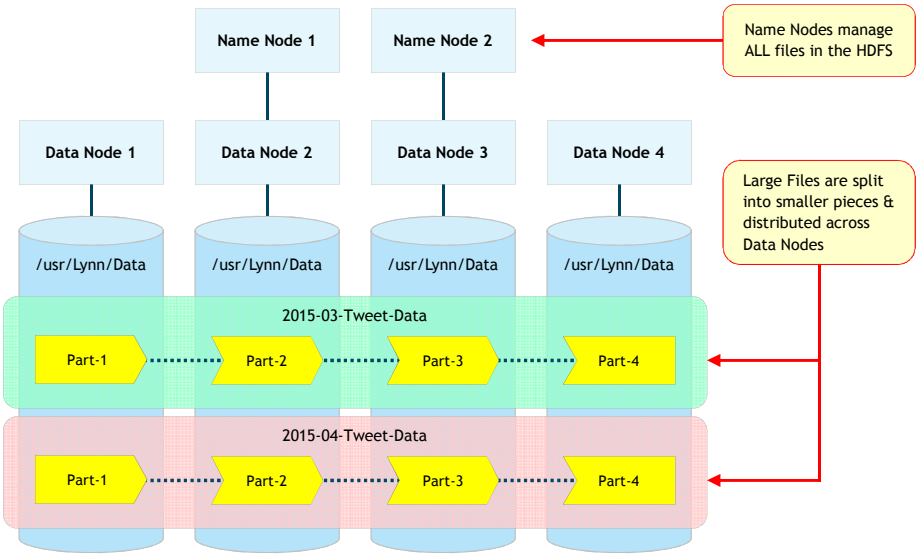
If I had 1000 clones, the work would finish 1000 times faster … … ***Linear Scalability***




## Large Files



## Hadoop File System



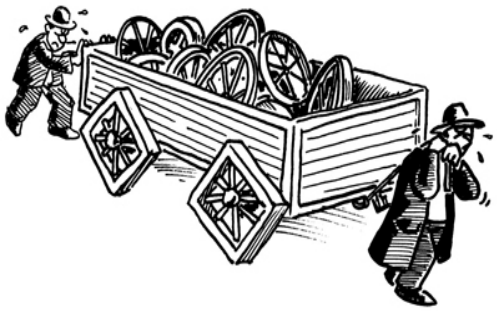




## The Right Tool for the Job

*"People shouldn't get ideas about turning off their relational systems and replacing them with Hadoop…*

*… As we start thinking about big data from the perspective of business needs, we're realizing that Hadoop isn't always the best tool for everything we need to do, and that using the wrong tool can sometimes be painful."*



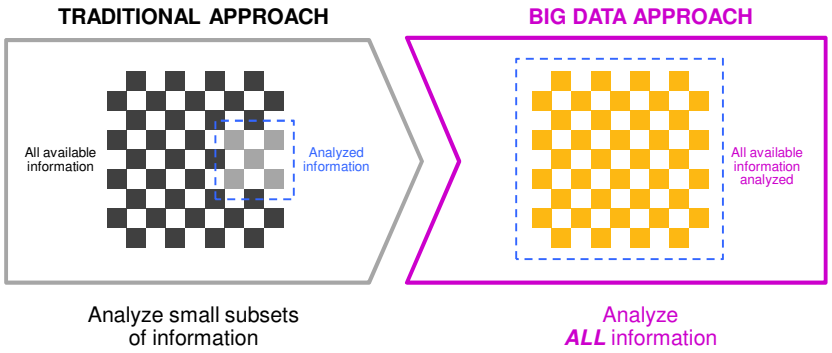
**Ken Rudin**
Head of Analytics at Facebook

## Big Data Methodologies

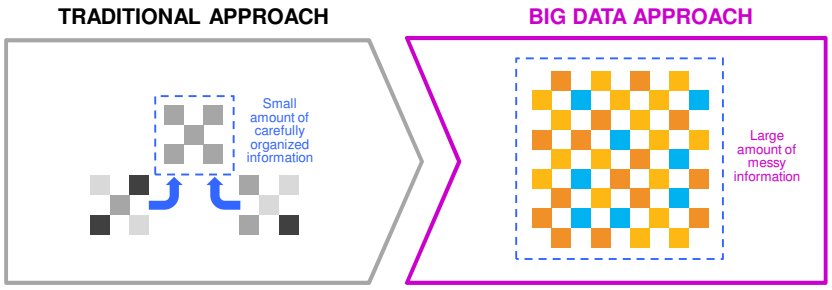
# Big Data Methodologies




## Leverage More of the Data Being Captured



## Reduce Effort Required to Leverage Data

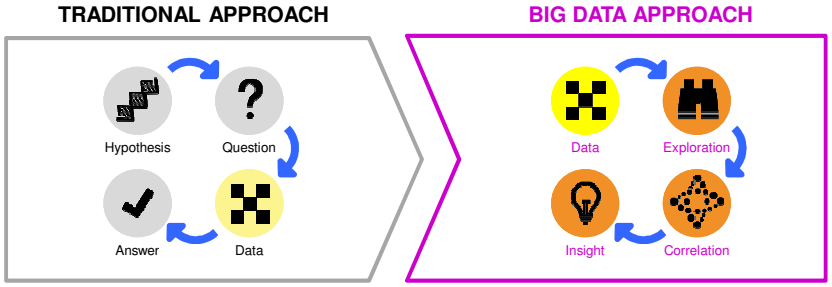



Carefully cleanse information ***before*** any analysis

Analyze information as is, cleanse as needed



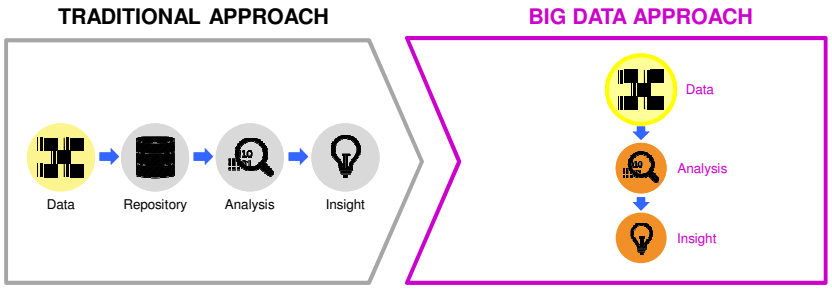

## Data Leads the Way




Start with hypothesis and test against selected data

Explore ***all*** data and identify correlations

## Leverage Data as it is Captured




Analyze data ***after*** it's been processed and landed in a warehouse or mart

Analyze data ***in motion*** as it's generated, in real-time




## NoSQL Databases

NoSQL Databases

## NoSQL Databases

- NoSQL: A term describing a class of Big Data Databases
- Its strange to define something by what it is NOT ?!?!?
- The term; *"NoSQL"* was created by Johan Oskarsson as a hash-tag to advertise an upcoming Meetup about BigData
- NoSQL does NOT mean NO SQL …
  … it means; "Not Only SQL."
- Characteristics of NoSQL
  - Non-Relational
  - Cluster Friendly
  - Schema-less
  - Designed to enable modern Web



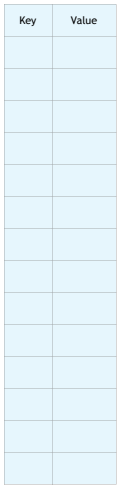

## Different Categories of noSQL Databases

| NoSQL Category | Use this when…. | Application Examples | Vendors |
|---|---|---|---|
| **Document** 63% revenue share* | • Schema is not well defined<br>• Schema is very likely to change, need to maintain flexibility<br>• Commonly described with JSON | • Frequently changing product catalogs | • Cloudant**<br>• MongoDB<br>• Couchbase<br>• MarkLogic |
| **Key-Value** 24% revenue share* | • Your data is not highly related<br>• All you need is basic Create, Read, Update, Delete (CRUD)<br>• Rapid Scaling for simple data collections | • User Sessions<br>• Shopping Cart | • Redis<br>• Aerospike<br>• AWS (DynamoDB)<br>• Basho Technologies (Riak) |
| **BigTable Columnar** 9% revenue share* | • High volume, low latency write<br>• Big Data, sparse data<br>• Need compression or versioning | • Telco, heavy ingest, petabyte scale<br>• User Activity logs<br>• Sensor data | • HBase (Hadoop)**<br>• BigTable<br>• Cassandra |
| **Graph DB** 4% revenue Share* | • Your data looks like a graph<br>• Have highly interconnected data, need to trace relationships | • Website Purchase Recommendations<br>• Social Network Processing | • Titan**<br>• Neo Technology (Neo4J) |

** Source: IBM study 2013 estimated by splitting total noSQL revenue ($288m) by ratio of top 10 vendors reported 2013 revenue. Total 2013 noSQL database revenue estimated $343m*
**** IBM Solutions of Choice.

## Key – Value Database

- Based on Amazon's Dynamo technology
- Data Model:
  - Key; Usually a scalar value
  - Value; Usually a scalar value
- Good for VERY Fast lookup
  - User-ID : Password
  - Account : Max Charge
  - SKU : Cost
- Examples: Dynomite, Voldemort, Tokyo

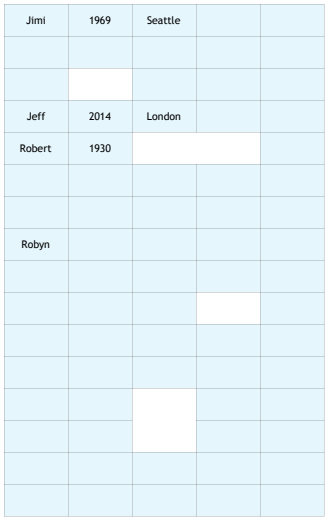






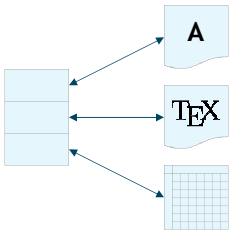
## Column Based DB

- Based on Google's BigTable technology
- Data Model
- Example: HBase, Hypertable, Cassandra



## Document Database

- Inspired by Lotus Notes
- Data model:
  - Collections of K-V collections
- Examples:
  - CouchDB
  - MongoDB
  - Riak

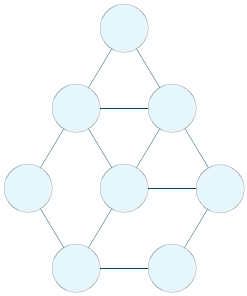




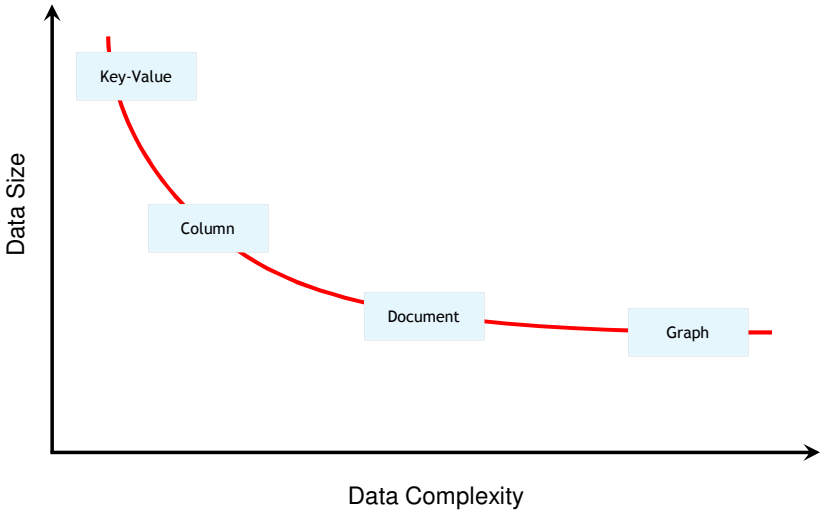


## Graph Database

- Derived from Euler & CS Graph Theory
- Data model:
  - Nodes,
  - Relations
  - K-V on both
- Examples: AllegroGraph, Sones, Neo4j



## NoSQL Data Models — Size ~ Complexity

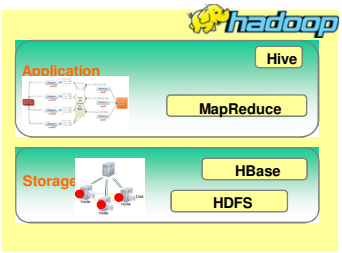






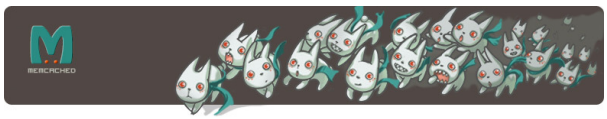
## NoSQL and Hadoop 101

- NoSQL is a generic term for a very nongeneric landscape of data stores
  - Evolved over time to serve a very different world of requirements
  - Social, blogs, clickstream, Cloud apps, mobile apps, complex relationships…
- Hadoop is a *framework* for processing large and varied data sets with low cost at a high degree of fault tolerance. Components include (among others):
  - A file system (Hadoop File System - HDFS) -
  - A framework for processing data (Map-Reduce)
  - May include a NoSQL Database - HBase

## The NoSQL Revolution

- Different requirements require different tools
  - Document stores
  - Key/value stores
  - BigTable implementations (columnar)
  - Graph databases
- Values (there are exceptions)
  - Huge data volumes – easy scale-out
  - Developers code integrity if it's needed
  - Relaxed (eventual) consistency
  - Semi-structured data
  - Schema on read






## Categories of NoSQL

| Type | Examples |
|---|---|
| Document store | CouchDB, mongoDB |
| Column store | Cassandra, HBase |
| Key-value store | Redis, riak |
| Graph store | InfiniteGraph, Neo4j |

Source: Akmal Chaudhri's NoSQLpresentation:

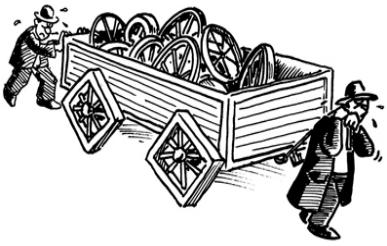
## Database Landscape Overview

| | SQL | noSQL database | Hadoop |
|---|---|---|---|
| **Description** | • Relational SQL (RDBMS)<br>• Operational and Analytic<br>• E.g. DB2, Oracle, Microsoft, Teradata, etc. | • noSQL database<br>• Mainly operational<br>• E.g. Cloudant, MongoDB, Redis, Riak, Aerospike, Amazon Dynamo DB, etc. | • SQL on Hadoop (mainly analytic)<br>• HBase (evolving OLTP, ACID)<br>• E.g. BigInsights, Cloudera, Hortonworks, MapR, Pivotal<br>• HP Labs Trafodion |
| **Typical Infrastructure** | • Proprietary database storage<br>• Unix, Linux, Windows<br>• SMP, MPP, SAN, Integrated Systems, Appliances | • Proprietary database storage<br>• Linux<br>• Commodity clusters<br>• Local attach disks, NAS<br>• Cloud<br>• Mobile | • HDFS files<br>• Linux<br>• Commodity clusters<br>• Local attach disks |
| **Primary Driver** | • Traditional I/T<br>• ACID | • Developer<br>• Agility, scalability, workload, cost | • Lower Cost<br>• All types of data |




## The Right Tool for the Job

**Big Data** ≠ **Hadoop**



*"There's a belief that if you want big data, you need to go out and buy Hadoop and then you're pretty much set. People shouldn't get ideas about turning off their relational systems and replacing them with Hadoop…*

*As we start thinking about big data from the perspective of business needs, we're realizing that Hadoop isn't always the best tool for everything we need to do, and that using the wrong tool can sometimes be painful."*

Ken Rudin
Head of Analytics at Facebook

# THANK YOU

"The art of progress is to preserve order amid change and to preserve change amid order"

Alfred North Whitehead