Big Data @ Yahoo

Matt Ahrens (mahrens@yahoo-inc.com)
Director of Engineering
Advertising Data & Analytics
My Story
Agenda

- Evolution of Big Data
  - Shift 1: The Rise of Hadoop (Scale)
  - Shift 2: The Need for Speed (Streaming)
  - Shift 3: The Opportunity for Learning (Science)

- Best Practices for Big Data

- Q & A
A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These titans—Alphabet (Google's parent company), Amazon, Apple, Facebook and
Agenda

- Evolution of Big Data
  - Shift 1: The Rise of Hadoop (Scale)
  - Shift 2: The Need for Speed (Streaming)
  - Shift 3: The Opportunity for Learning (Science)

- Best Practices for Big Data

- Q & A
Big Data Investment

- Data keeps growing

**Cloud Capital**

The three giants of cloud infrastructure are spending lavishly to keep up with one another, and distance themselves from rivals.

**Capital expenses, in billions**

- Amazon: $4.0B
- Alphabet: $2.9B
- Microsoft: $2.0B

Q4 2016

*Amazon’s spending includes property and equipment acquired under capital leases, a key way the company supplements its capital spending.

Source: the companies

THE WALL STREET JOURNAL.
Relational Databases -- limitations

- In early days of web, relational databases were sufficient for storing web logs

- Transactions would be stored and clusters of databases would scale as needed

- Limitations
  - Defined schema -- need to know data format
  - Scale overhead -- procure and set up new hardware
  - Scale ceiling -- up to GBs, but TBs/PBs not feasible or cost-effective
The Past Architecture

Batch data input

Custom Cluster #1
Transforms

Custom Cluster #2
Joins

Custom Cluster #3
Validation

Custom Cluster #4
Aggregations

Data Warehouse (Custom Format)

SQL Layer

Proxy Server

Data Users / Customers
The Elephant Comes Into The Room
Why Move To Hadoop?

- Legacy systems were not performing well (< 1 TB / day)
- We had customers who wanted access to raw feeds (TB/day per customer)
- The advertising roadmap called for a 5-10x increase in traffic (new features, new customers onboarding)

<table>
<thead>
<tr>
<th>Year</th>
<th>$/GB Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>$437,500.00</td>
</tr>
<tr>
<td>1985</td>
<td>$105,000.00</td>
</tr>
<tr>
<td>1990</td>
<td>$11,200.00</td>
</tr>
<tr>
<td>1995</td>
<td>$1,120.00</td>
</tr>
<tr>
<td>2000</td>
<td>$11.00</td>
</tr>
<tr>
<td>2005</td>
<td>$1.24</td>
</tr>
<tr>
<td>2010</td>
<td>$0.09</td>
</tr>
<tr>
<td>2013</td>
<td>$0.05</td>
</tr>
<tr>
<td>2014</td>
<td>$0.03</td>
</tr>
</tbody>
</table>

Source: www.statisticbrain.org
The Architecture on Hadoop

Hadoop
- Map-Reduce
- Pig
- Hive
- Oozie

Access
- User groups
- Easy onboard

Scale
- 45 days raw data
- Full event logs

Transforms

Joins

Validation

Aggregations

Batch data input

HDFS

Proxy Server

Data Users / Customers
Agenda

- Evolution of Big Data
  - Shift 1: The Rise of Hadoop (Scale)
  - Shift 2: The Need for Speed (Streaming)
  - Shift 3: The Opportunity for Learning (Science)

- Best Practices for Big Data

- Q & A
How Did We Get Here?

- People always have wanted data faster.
- Finally we had hardware costs that were in line with doing in-memory streaming for billions of events/day.

<table>
<thead>
<tr>
<th>Year</th>
<th>$/GB RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>$6,328,125.00</td>
</tr>
<tr>
<td>1985</td>
<td>$859,375.00</td>
</tr>
<tr>
<td>1990</td>
<td>$103,880.00</td>
</tr>
<tr>
<td>1995</td>
<td>$30,875.00</td>
</tr>
<tr>
<td>2000</td>
<td>$1,107.00</td>
</tr>
<tr>
<td>2005</td>
<td>$189.00</td>
</tr>
<tr>
<td>2010</td>
<td>$12.37</td>
</tr>
<tr>
<td>2013</td>
<td>$5.50</td>
</tr>
</tbody>
</table>

Source: www.statisticbrain.org
The Lambda Architecture: Real-Time + Batch
The Present Architecture

- **Hadoop**
  - Transforms
  - Joins
  - Validation
  - Aggs

- **Storm**
  - Spout
  - Bolt
  - Bolt
  - Sink

**Batch data input** via HDFS

**Real-time data input** via Storm

- **Druid**

- **Data Users / Customers**
In-Memory Distributed Query Databases

- Druid (open source)
- Redshift (Amazon)
- Impala (Cloudera, open source)
- Presto (Facebook, open source)
- Hive ORC (Yahoo/HortonWorks, open source)
Agenda

■ Evolution of Big Data
  ● Shift 1: The Rise of Hadoop (Scale)
  ● Shift 2: The Need for Speed (Streaming)
  ● Shift 3: The Opportunity for Learning (Science)

■ Best Practices for Big Data

■ Q & A
This is your machine learning system?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.
The Opportunity For Learning
Data Analytics Landscape

■ Past
  ● Descriptive Analytics
    ● *What happened?*
  ● Diagnostic Analytics
    ● *Why did it happen?*

■ Future
  ● Predictive Analytics
    ● *What is going to happen?*
  ● Prescriptive Analytics
    ● *How do we impact what is going to happen?*
Machine Learning @ Scale

- With the rise of big data has come the application of various machine learning techniques at scale.

- Frameworks have followed: Spark, TensorFlow, Pandas, and more.

- Desire to go beyond past analytics (what happened and why) to future analytics (what is going to happen and how can we change what’s going to happen).
Obstacles for Machine Learning @ Scale

- Data size
  - Storing TBs of data in memory for iterative processing can be costly (requires RAM investment)
  - Hyper-tuning and model selection can take days/weeks

- Query latency
  - TB queries can take minutes, PB queries can take hours

- Fragmented frameworks and libraries
The Data Lake

Data Lake

- Sensoren
- Web-Logs
- Social Media
- Mobile
- Search
- Marketing
- Transactions
- Demographics
- Data Warehouse
- CRM

From pmone.com
Disk Access Latency: The Last Frontier

From [https://maxkanaskar.files.wordpress.com/](https://maxkanaskar.files.wordpress.com/)
The Dream: An Interactive Data Lake

**Vision:** interactive (sub-second) query capabilities for PBs data

![Diagram](image)

- **Real-time data input**
- **Storm**
  - Spout
  - Bolt
  - Bolt
  - Sink

**Data Scientists**
- Machine Learning frameworks and libraries compatible with Data Lake

**Data Lake** (PBs of raw data)

**Applications**

**Business Users**
- Standard SQL interface with visualizations available for sharing
Agenda

- Evolution of Big Data
  - Shift 1: The Rise of Hadoop (Scale)
  - Shift 2: The Need for Speed (Streaming)
  - Shift 3: The Opportunity for Learning (Science)

- Best Practices for Big Data

- Q & A
Build For Open Access
Build For Open Access

- Democratize data by choosing an appropriate tech stack

- Questions to consider in technology choice
  - What is the onboarding process for new users?
  - What technical knowledge or skillset is needed to use the data?
  - How well does the technology interface with other systems in use or planned to be used?
Govern The Data
Discover
- Data discovery
- Data profiling
- Data inventories
- Process inventories
- CRUD analysis
- Capabilities assessment

Define
- Business glossary creation
- Data classifications
- Data relationships
- Reference data
- Business rules
- Data governance policies
- Other dependent policies
- Key Performance Indicators

Measure and Monitor
- Proactive monitoring
- Operational dashboards
  - Reactive operational DQ audits
  - Dashboard monitoring/audits
- Data lineage analysis
- Program performance
- Business value/ROI

Apply
- Automated rules
- Manual rules
- End to end workflows
- Business/IT collaboration

Data Governance
Why Data Governance Is Needed

- Lack of standards and oversight creates friction
  - People can’t find data
  - People use data for the wrong use case
  - Data is not clean or is incomplete

- Treat internal data consumers as external customers

- Tips
  - Directory -- list of location/format for datasets
  - Dictionary -- what, how, when for each dataset
Innovate With Data
Innovate With Data

- Allocate time and resources to allow for data exploration and innovation

Benefits
- Better understanding of what is in the data
- More quickly detect data quality issues
- Cross-organization data use cases arise

Tips
- Keep a backlog of data exploration ideas
- Hold a data hack day to encourage innovation
Visualize For Impact
StackOverflow.com gets 23% of its users from the US and its traffic dips on the weekend.

From quantcast.com
StackOverflow.com users are mostly Male, make over $150K, are between 18-24, and have grad school education.

From quantcast.com
Visualize For Impact

■ When sharing insights derived from data, graphics will be more impactful than text

■ Consider what main effect you want from your data and choose a visualization accordingly

■ Build a data visualization toolkit -- leverage existing libraries in R, Python, Javascript
Agenda

■ Evolution of Big Data
  ● Shift 1: The Rise of Hadoop (Scale)
  ● Shift 2: The Need for Speed (Streaming)
  ● Shift 3: The Opportunity for Learning (Science)

■ Best Practices for Big Data

■ Q & A
Q & A