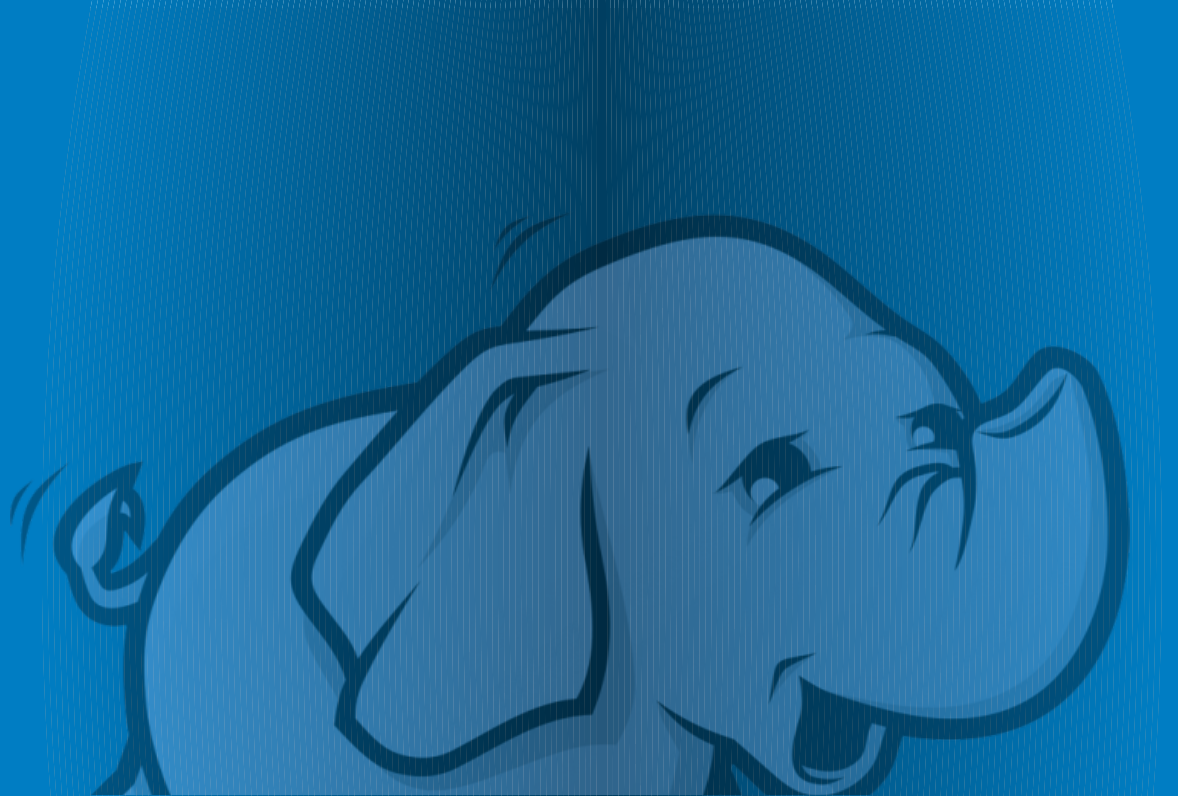


October 2011 – CHADNUG – Chattanooga, TN

What is Hadoop?

Josh Patterson | Sr Solution Architect

cloudera

cloudera

Who is Josh Patterson?

- josh@cloudera.com
  - Twitter: @jpatanooga
- Master's Thesis: self-organizing mesh networks
  - Published in IAAI-09: TinyTermite: A Secure Routing Algorithm
- Conceived, built, and led Hadoop integration for openPDC project at Tennessee Valley Authority (TVA)
  - Led team which designed classification techniques for time series and Map Reduce
- Open source work at
  - http://openpdc.codeplex.com
  - https://github.com/jpatanooga
- Today
  - Sr. Solutions Architect at Cloudera

Outline

- Let's Set the Stage
- Story Time: Hadoop and the Smartgrid
- The Enterprise and Hadoop
- Use Cases and Tools

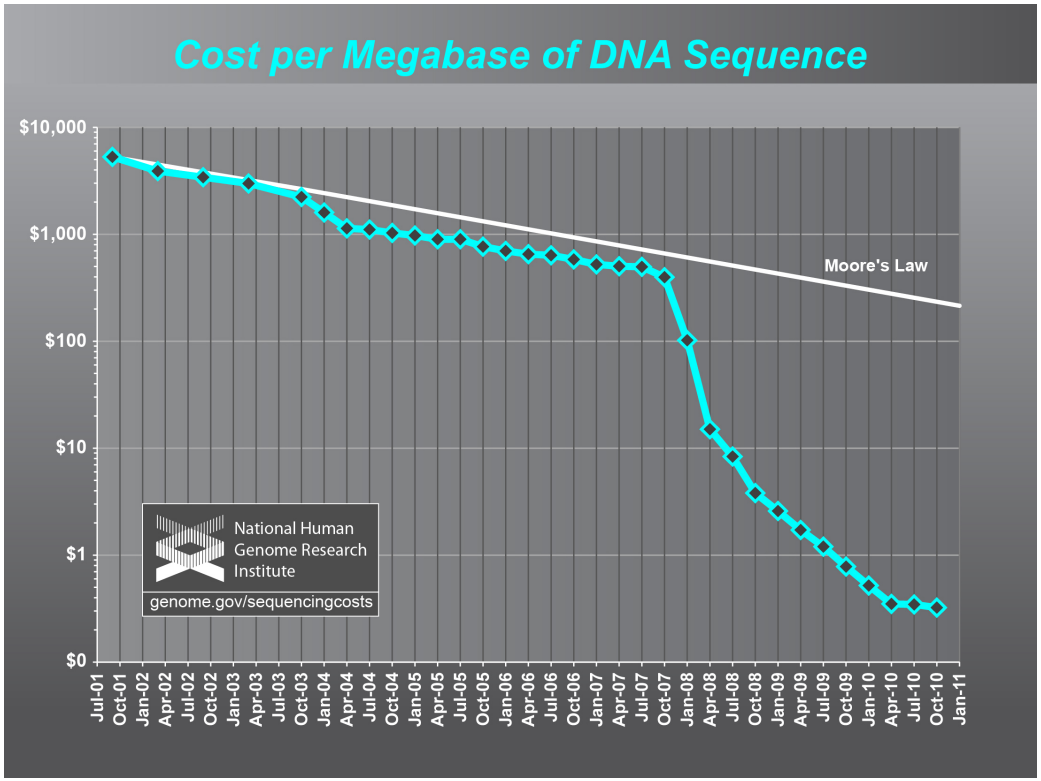
"After the refining process, one barrel of crude oil yielded more than 40% gasoline and only 3% kerosene, creating large quantities of waste gasoline for disposal."

--- Excerpt from the book "The American Gas Station"

Data Today: The Oil Industry Circa 1900

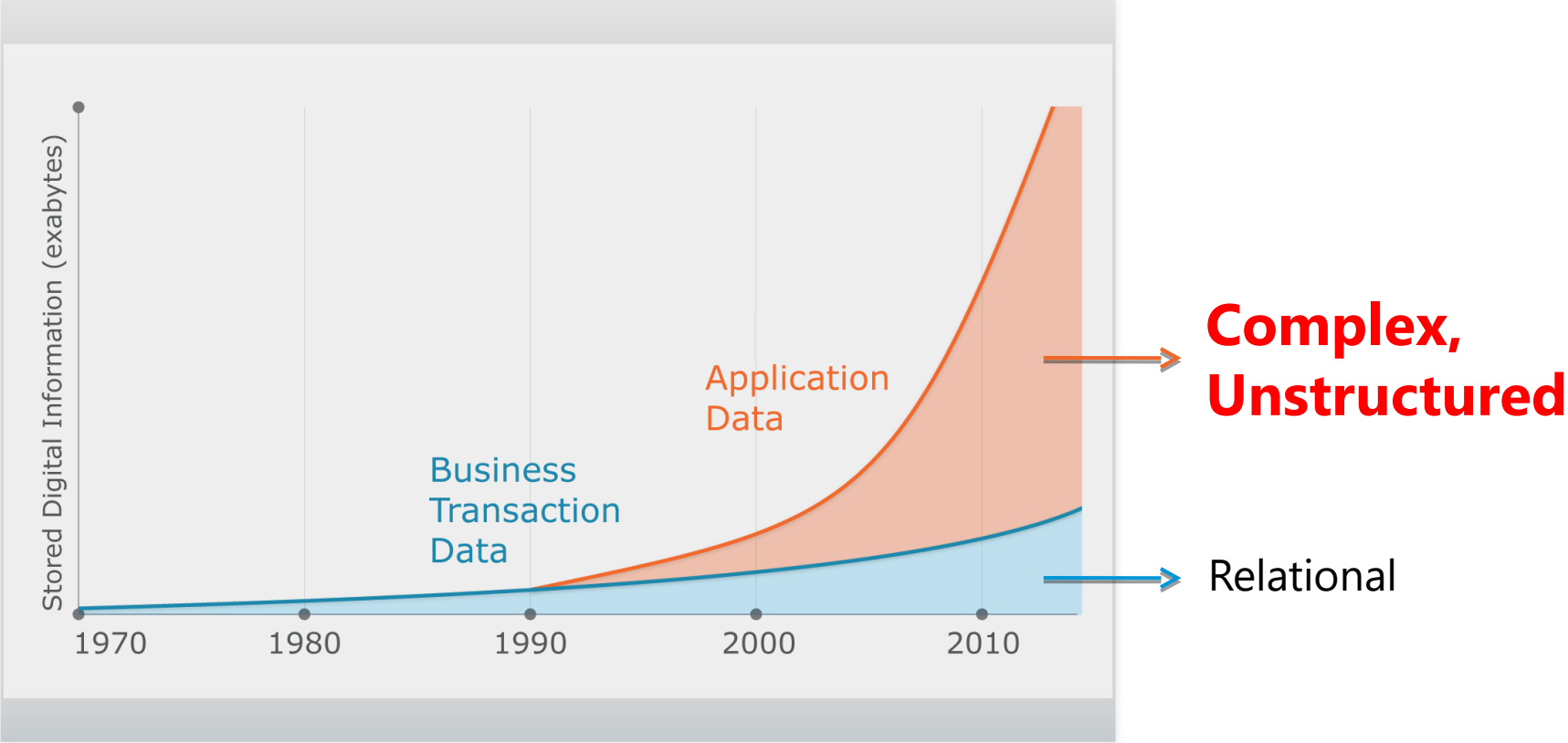
DNA Sequencing Trends

- Cost of DNA Sequencing **Falling Very Fast**



cloudera

Unstructured Data Explosion



- 2,500 exabytes of new information in 2012 with Internet as primary driver
- Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 “zettabytes” this year

cloudera

Obstacles to Leveraging Data

- Data comes in many shapes and sizes: relational tuples, log files, semistructured textual data (e.g., e-mail)
  - Sometimes makes the data **unwieldy**
- Customers are **not creating schemas** for all of their data
  - Yet still may want to join data sets
- Customers are moving some of it to **tape** or **cold storage**, **throwing it away** because "it doesn't fit"
  - They are **throwing data** away **because its too expensive to hold**
  - ***Similar to the oil industry in 1900***

cloudera

A Need for a Platform in an Evolving Landscape

- Need ability to look at true distribution of data
  - Previously impossible due to scale
- Need lower cost of analysis
  - Ad Hoc analysis now more open and flexible
- Need Greater Flexibility, "**BI Agility**"
  - Less restrictive than SQL-only systems
- **Speed @ Scale is the new Killer App**
  - Results in that previously took 1 day to process can gain new value when created in 10 minutes.

cloudera

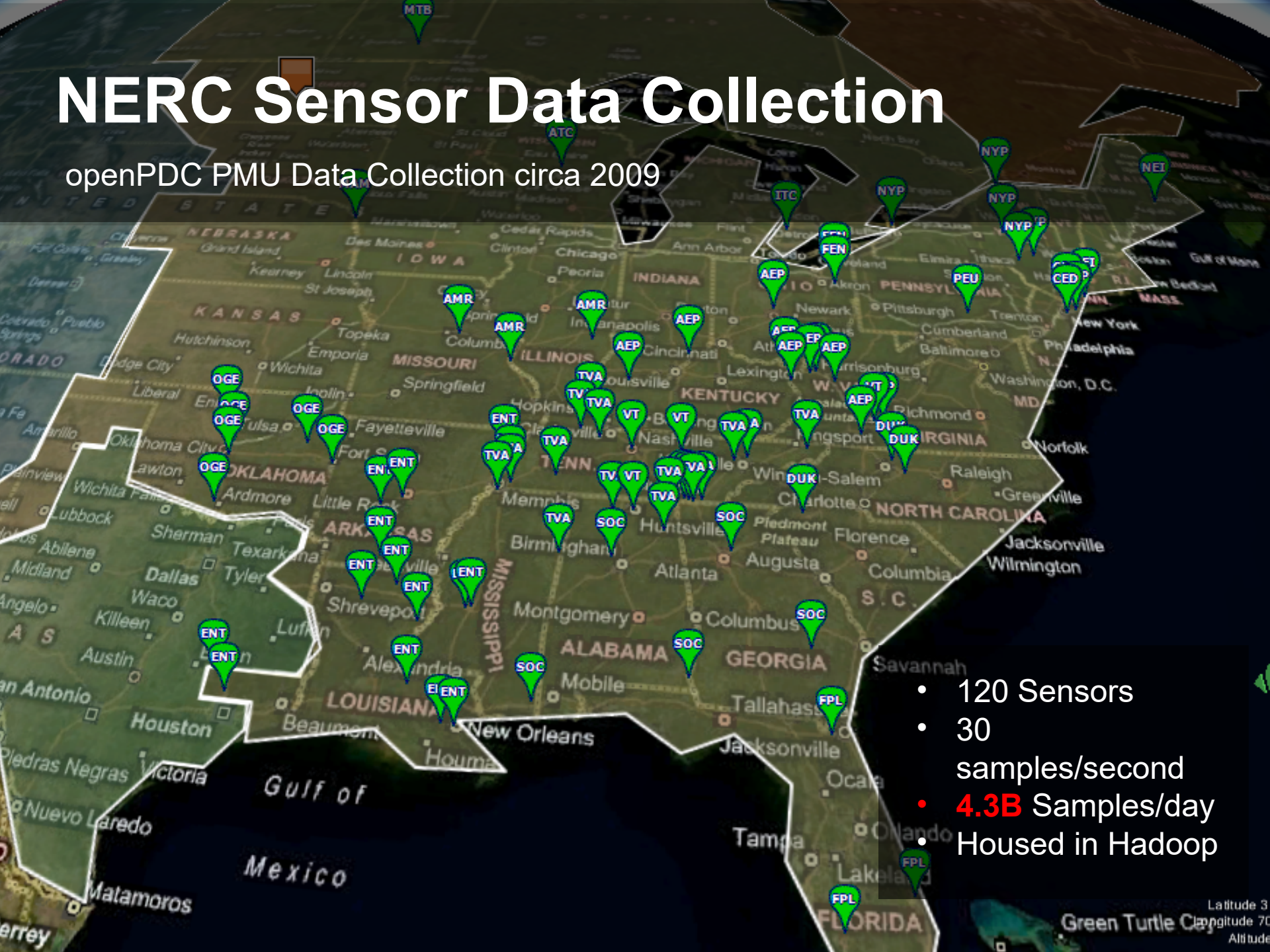
"We're gonna need a bigger boat."

--- Roy Scheider, "Jaws"

Story Time: Hadoop and the Smartgrid

NERC Sensor Data Collection

openPDC PMU Data Collection circa 2009



- 120 Sensors
- 30 samples/second
- **4.3B** Samples/day
- Housed in Hadoop

NERC Wanted High-Res Smartgrid Data

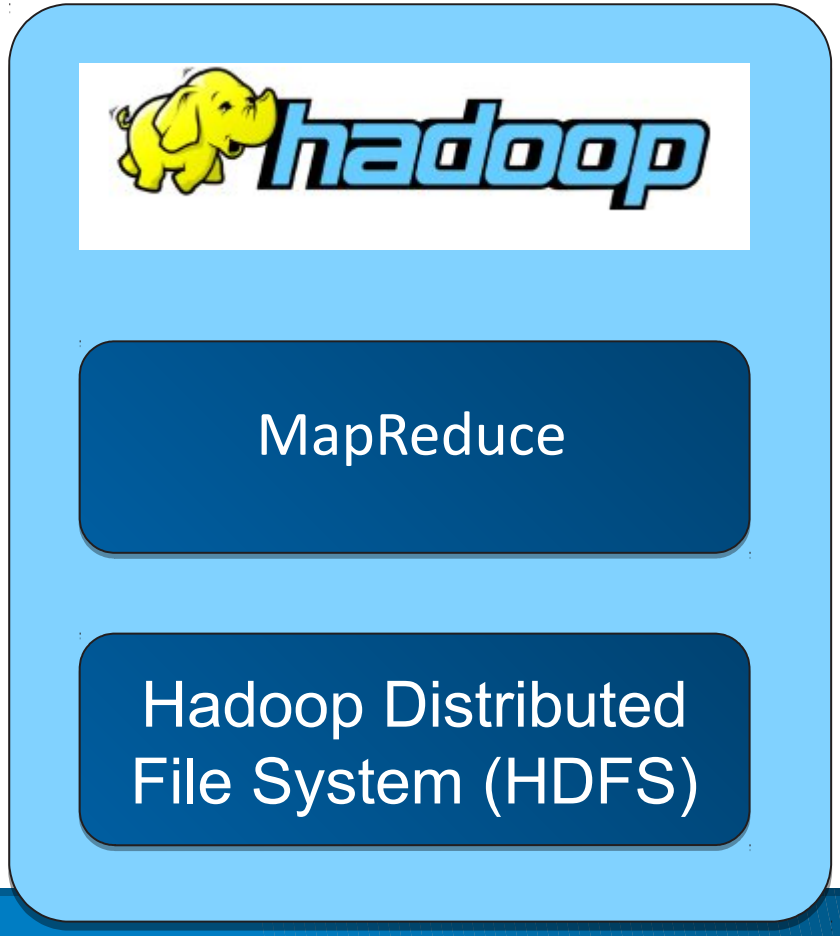
- Started **openPDC** project @ TVA
  - http://openpdc.codeplex.com/
- We used **Hadoop** to store and process **smartgrid** (PMU) time series data
  - https://openpdc.svn.codeplex.com/svn/Hadoop/Current%20Version/

Major Themes From openPDC

- Velocity of incoming data
- Where to put the data?
  - Wanted to scale out, not up
  - Wanted **linear scalability** in **cost vs size**
  - Wanted system **robust** in the face of **HW failure**
  - **Not fans of vendor lock-in**
- What can we realistically expect from analysis and extraction at this scale?
  - How long does it take to scan a Petabyte @ 40MB/s?

Apache Hadoop

Open Source Distributed Storage and Processing Engine



- **Consolidates Mixed Data**
  - Move complex and relational data into a single repository
- **Stores Inexpensively**
  - Keep raw data always available
  - Use industry standard hardware
- **Processes at the Source**
  - Eliminate ETL bottlenecks
  - Mine data first, govern later

cloudera

What Hadoop does

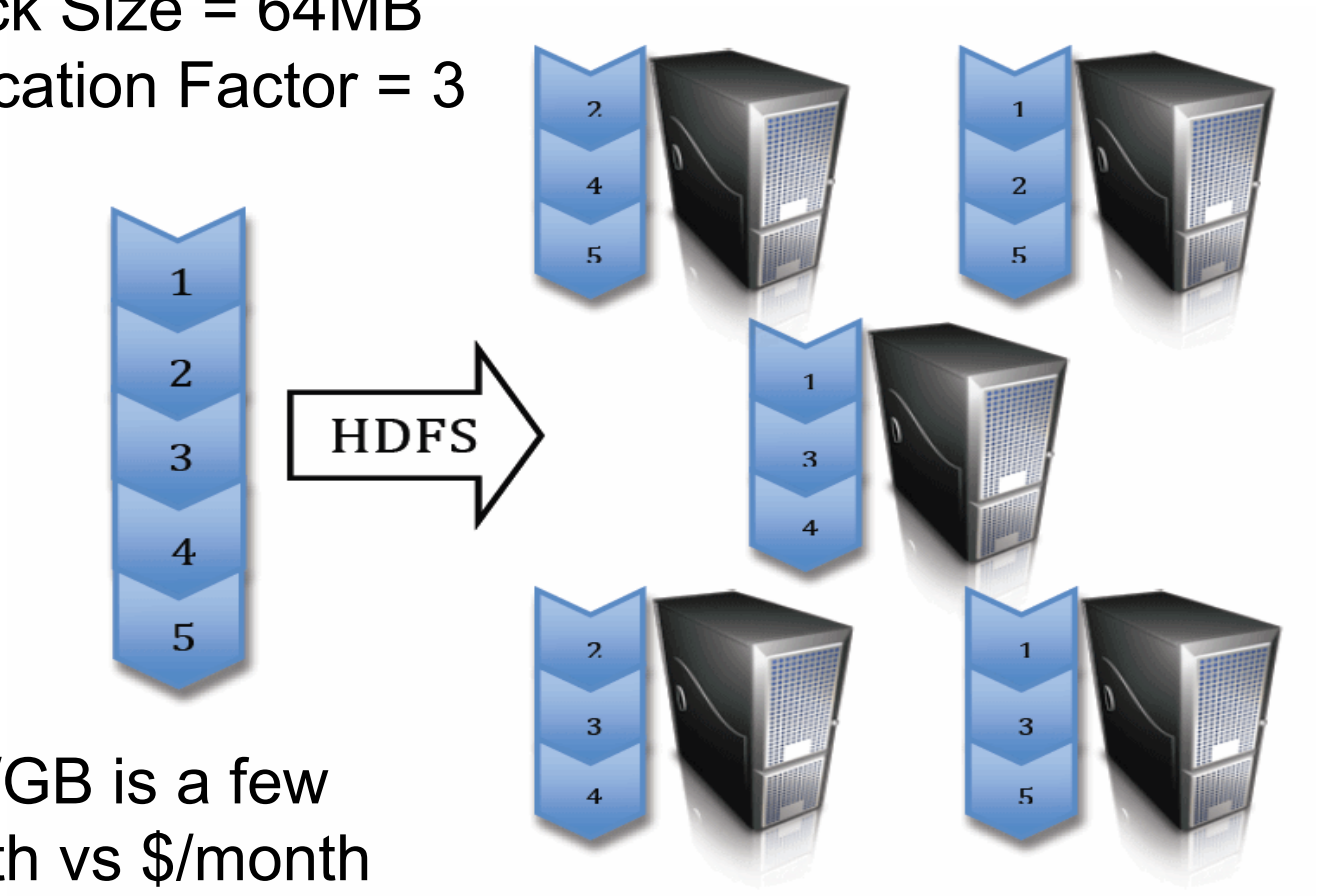
- Networks industry standard hardware nodes together to combine compute and storage into scalable distributed system
- Scales to petabytes without modification
- Manages fault tolerance and data replication automatically
- Processes semi-structured and unstructured data easily
- Supports MapReduce natively to analyze data in parallel

What Hadoop does not do

- No random access or transactions
  - **Hadoop is not a database**
- Not real-time
  - **Hadoop is batch oriented**
- Push-button install or configure
  - **Hadoop requires some advanced skills**
  - This aspect is part of Cloudera's value add

HDFS: Hadoop Distributed File System

Block Size = 64MB
Replication Factor = 3



Cost/GB is a few ¢/month vs $/month

MapReduce

- In simple terms, it's an application with 2 functions
  - Map Function
    - Think massively parallel "**group by**"
  - Reduce Function
    - Think "**aggregation + processing**"
- Not hard to write
  - Can be challenging to refactor existing algorithms
- Designed to work hand-in-hand with HDFS
  - Minimizes disk seeks, operates at "transfer rate" of disk

Speed @ Scale

- Scenario
  - 1 million sensors, collecting sample / 5 min
  - 5 year retention policy
  - Storage needs of 15 TB
- Processing
  - Single Machine: 15TB takes 2.2 DAYS to scan
  - MapReduce@ 20 nodes: **Same task takes 11 Minutes**

MapReduce Tools

- Pig
  - Procedural language compiled into MR
- Hive
  - **SQL-like** language compiled into MR
- Mahout
  - Collection of data mining algorithms for Hadoop
- Streaming
  - Ability to write MR with tools such as python, etc

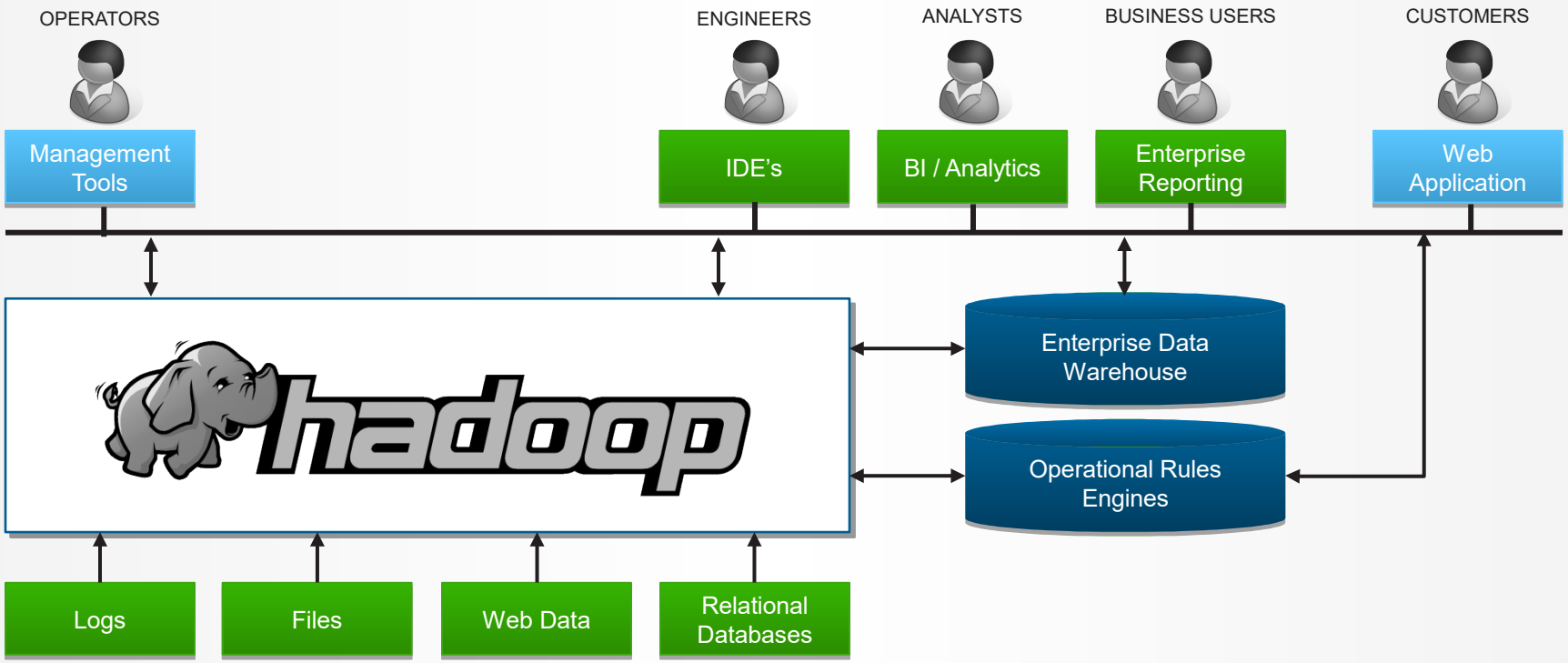
“What if you don’t know the questions?”

--- Forrester Report

The Enterprise and Hadoop

Apache Hadoop in Production

How Apache Hadoop fits
into your existing infrastructure.



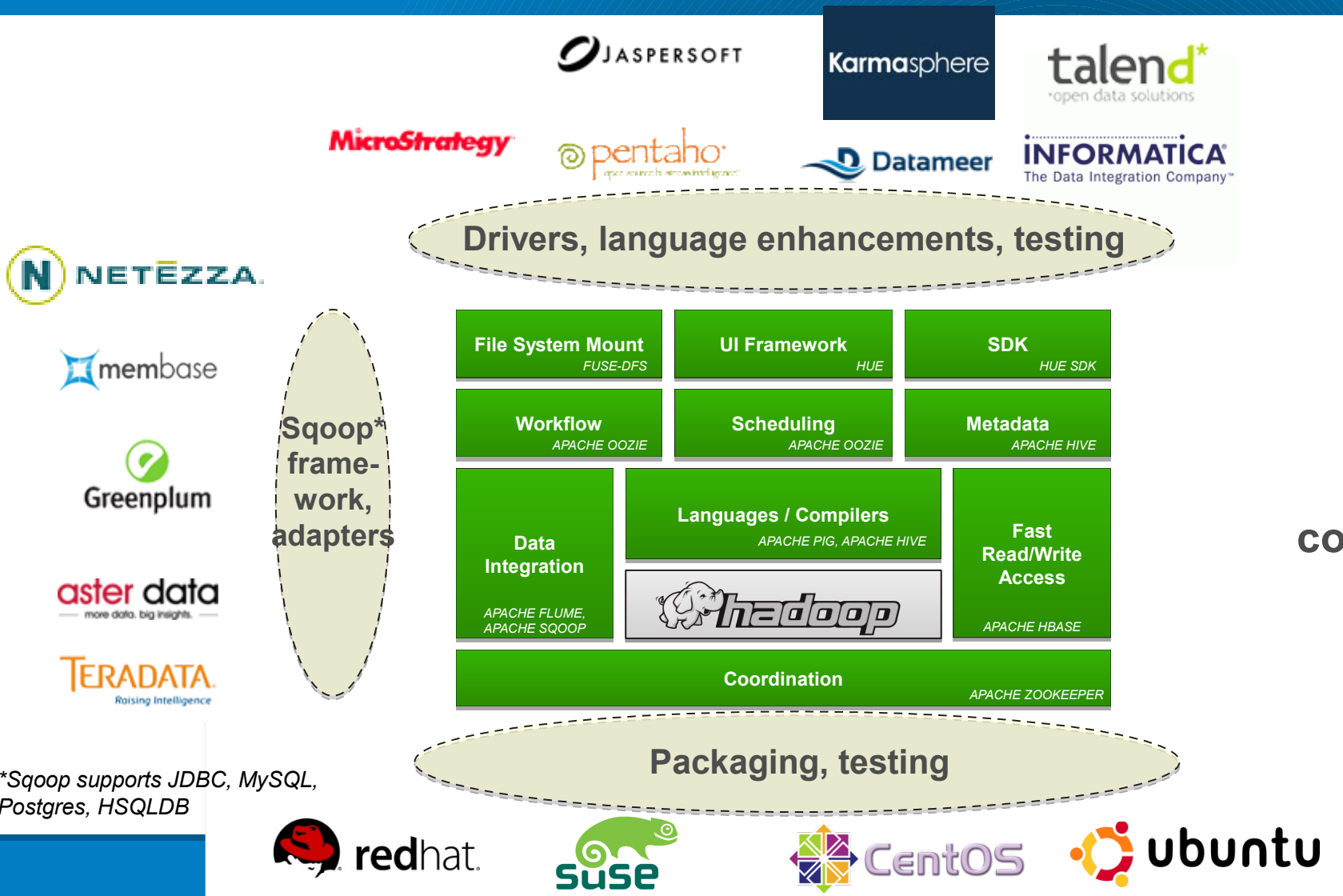
cloudera

What the Industry is Doing

- Microsoft
  - ships CTP of **Hadoop Connectors for SQL Server and Parallel Data Warehouse**
    - (based on Cloudera's Sqoop)
  - http://blogs.technet.com/b/dataplatforminsider/archive/2011/08/25/microsoft-ships-ctp-of-hadoop-connectors-for-sql-server-and-parallel-data-warehouse.aspx
- Oracle
  - Announcments at Oracle Open World
    - Connector to **Hadoop** allowing data flow into Oracle
    - **Hadoop** Accelerator for Exalogic
    - In-memory processing for MapReduce
    - ETL using **Hadoop**
    - Integrated analytics on Oracle and **Hadoop**

cloudera

Integrating With the Enterprise IT Ecosystem



More coming…

*Sqoop supports JDBC, MySQL, Postgres, HSQLDB

Forester Report

- World Economic Forum
  - **Declared that data is a new asset class**
- Big data is an **applied science project** in most companies
  - Major potential constraint is **not** the cost of the computing technology
  - **but** the **skilled people** needed to carry out these projects
    - the data scientists
- What if you don't know the questions?
  - Big data is all about exploration without preconceived notions
  - Need tools to ask questions to understand the right questions to ask
- Much of the software is based on open-source **Hadoop**
  - http://blogs.forrester.com/brian_hopkins/11-09-30-big_data_will_help_shape_your_markets_next_big_winners

cloudera

Ever been recommended a friend on Facebook?

Ever been recommended a product on Amazon?

Ever used the homepage at Yahoo?

What Can Hadoop Do For Me?

Problems Addressed With Hadoop

- Text Mining
- Index Building
  - Search Engines
- Graph Creation
  - Twitter, Facebook
- Pattern Recognition
  - Naïve Bayes Classification
- Recommendation Engines
- Predictive Models
- Risk Assessment

A Few Named Examples

Analyze search terms and subsequent user purchase decisions to tune search results, increase conversion rates

Digest long-term historical trade data to identify fraudulent activity and build real-time fraud prevention

Model site visitor behavior with analytics that deliver better recommendations for new purchases

Continually refine predictive models for advertising response rates to deliver more precisely targeted advertisements

Replace expensive legacy ETL system with more flexible, cheaper infrastructure that is 20 times faster

Correlate educational outcomes with programs and student histories to improve results

cloudera

Packages For Hadoop

- DataFu
  - From Linkedin
    - http://sna-projects.com/datafu/
  - UDFs in Pig
  - used at LinkedIn in many of off-line workflows for data derived products
    - "People You May Know”
    - "Skills”
  - Techniques
    - PageRank
    - Quantiles (median), variance, etc.
    - Sessionization
    - Convenience bag functions
    - Convenience utility functions

Integration with Libs

- Mix MapReduce with Machine Learning Libs
  - WEKA
  - KXEN
  - CPLEX
- Map side “groups data”
- Reduce side processes groups of data with Lib in parallel
  - Involves tricks in getting K/V pairs into lib
  - Pipes, tmp files, task cache dir, etc

Ask the right questions up front..

- Is the job disk bound?
- What is the latency requirement on the job?
  - Does it need a sub-minute latency? (not good for Hadoop!)
- Does the job look at a lot or all of the data at the same time?
  - Hadoop is good at looking at all data, complex/fuzzy joins
- Is large amounts of ETL processing needed before analysis?
  - Hadoop is good at ETL pre-processing work
- Can the analysis be converted into MR / Pig / Hive?

cloudera

What Hadoop Not Good At in Data Mining

- Anything highly iterative
- Anything that is extemely CPU bound and not disk bound
- Algorithms that can't be inherently parallelized
  - Examples
    - Stochastic Gradient Descent (SGD)
    - Support Vector Machines (SVM)
  - Doesn't mean they arent great to use

Questions? (Thank You!)

- **Hadoop World 2011**
  - http://www.hadoopworld.com/
- Cloudera's Distribution including Apache Hadoop (CDH):
  - http://www.cloudera.com
- Resources
  - http://www.slideshare.net/cloudera/hadoop-as-the-platform-for-the-smartgrid-at-tva
  - http://gigaom.com/cleantech/the-google-android-of-the-smart-grid-openpdc/
  - http://news.cnet.com/8301-13846_3-10393259-62.html
  - http://gigaom.com/cleantech/how-to-use-open-source-hadoop-for-the-smart-grid/
- Timeseries blog article
  - http://www.cloudera.com/blog/2011/03/simple-moving-average-secondary-sort-and-mapreduce-part-1/

More?

- Look at www.cloudera.com/training to learn more about Hadoop
- Read www.cloudera.com/blog
  - Lots of great use cases.
- Check out the downloads page at
  - www.cloudera.com/downloads
  - Get your own copy of Cloudera Distribution for Apache Hadoop (CDH)
  - Grab Demo VMs, Connectors, other useful tools.

- Contact Josh with any questions at
  - **josh@cloudera.com**

cloudera

October 2011 – CHADNUG – Chattanooga, TN

What is Hadoop?

Josh Patterson | Sr Solution Architect

cloudera

cloudera

Its all about the love, baby.

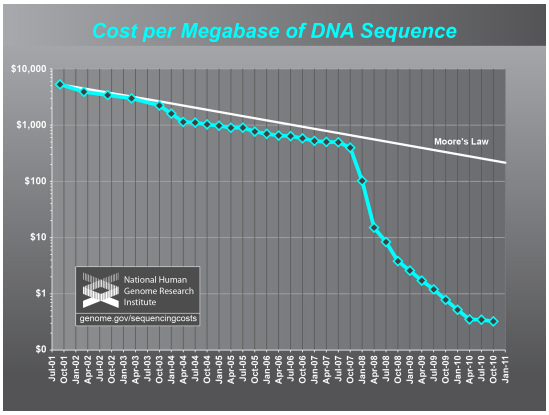
Outline

- Let's Set the Stage
- Story Time: Hadoop and the Smartgrid
- The Enterprise and Hadoop
- Use Cases and Tools

Theme: they through away a lot of valuable gas and oil just like we through away data today

DNA Sequencing Trends

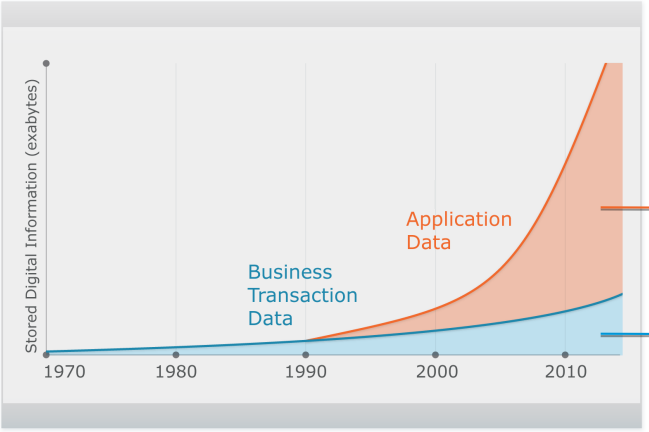
- Cost of DNA Sequencing **Falling Very Fast**



cloudera

Example of data production trends

Unstructured Data Explosion



Complex, Unstructured

Relational

- 2,500 exabytes of new information in 2012 with Internet as primary driver
- Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 "zettabytes" this year

cloudera

Obstacles to Leveraging Data

- Data comes in many shapes and sizes: relational tuples, log files, semistructured textual data (e.g., e-mail)
  - Sometimes makes the data **unwieldy**
- Customers are **not creating schemas** for all of their data
  - Yet still may want to join data sets
- Customers are moving some of it to **tape** or **cold storage**, **throwing it away** because "it doesn't fit"
  - They are **throwing data** away **because its too expensive to hold**
  - ***Similar to the oil industry in 1900***

cloudera

But what if some constraints changed?

A Need for a Platform in an Evolving Landscape

- Need ability to look at true distribution of data
  - Previously impossible due to scale
- Need lower cost of analysis
  - Ad Hoc analysis now more open and flexible
- Need Greater Flexibility, "**BI Agility**"
  - Less restrictive than SQL-only systems
- **Speed @ Scale is the new Killer App**
  - Results in that previously took 1 day to process can gain new value when created in 10 minutes.

cloudera

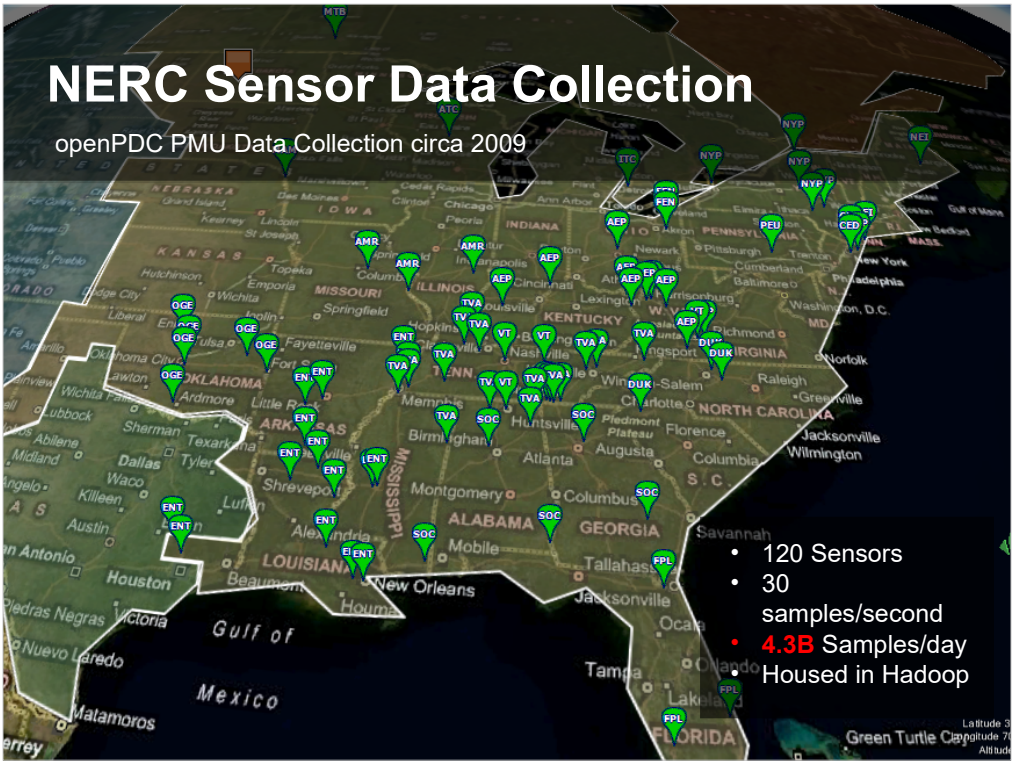
Talk about changing market dynamics of storage cost

What if some of the previously held constraints changed? Enter hadoop

“We’re gonna need a bigger boat.”

--- Roy Scheider, “Jaws”

Story Time: Hadoop and the Smartgrid



Let’s set the stage in the context of story, why we were looking at big data for time series.

NERC Wanted High-Res Smartgrid Data

- Started **openPDC** project @ TVA
  - http://openpdc.codeplex.com/
- We used **Hadoop** to store and process **smartgrid** (PMU) time series data
  - https://openpdc.svn.codeplex.com/svn/Hadoop/Current%20Version/

cloudera

Ok, so how did we get to this point?

Older SCADA systems take 1 data point per 2-4 seconds --- PMUs --- 30 times a sec, 120 PMUs, Growing by 10x factor

Major Themes From openPDC

- Velocity of incoming data
- Where to put the data?
  - Wanted to scale out, not up
  - Wanted **linear scalability** in **cost vs size**
  - Wanted system **robust** in the face of **HW failure**
  - **Not fans of vendor lock-in**
- What can we realistically expect from analysis and extraction at this scale?
  - How long does it take to scan a Petabyte @ 40MB/s?

cloudera

Data was sampled 30 times a second

Number of sensors (Phasor Measurement Units / PMU) was increasing rapidly

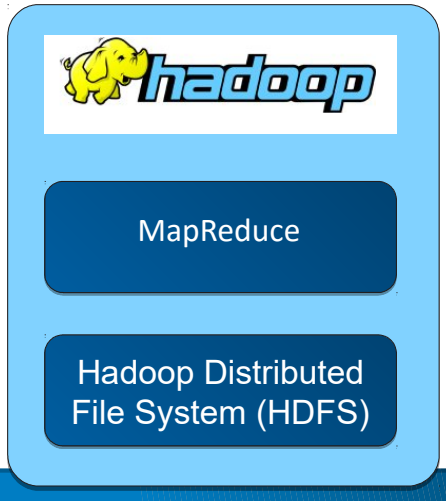
was 120, heading towards **1000** over next 2 years, currently taking in **4.3 billion** samples per day

Cost of SAN storage became excessive

Little analysis possible on SAN due to poor read rates on large amounts (TBs) of data

Apache Hadoop

Open Source Distributed Storage and Processing Engine



- **Consolidates Mixed Data**
  - Move complex and relational data into a single repository
- **Stores Inexpensively**
  - Keep raw data always available
  - Use industry standard hardware
- **Processes at the Source**
  - Eliminate ETL bottlenecks
  - Mine data first, govern later

cloudera

What Hadoop does

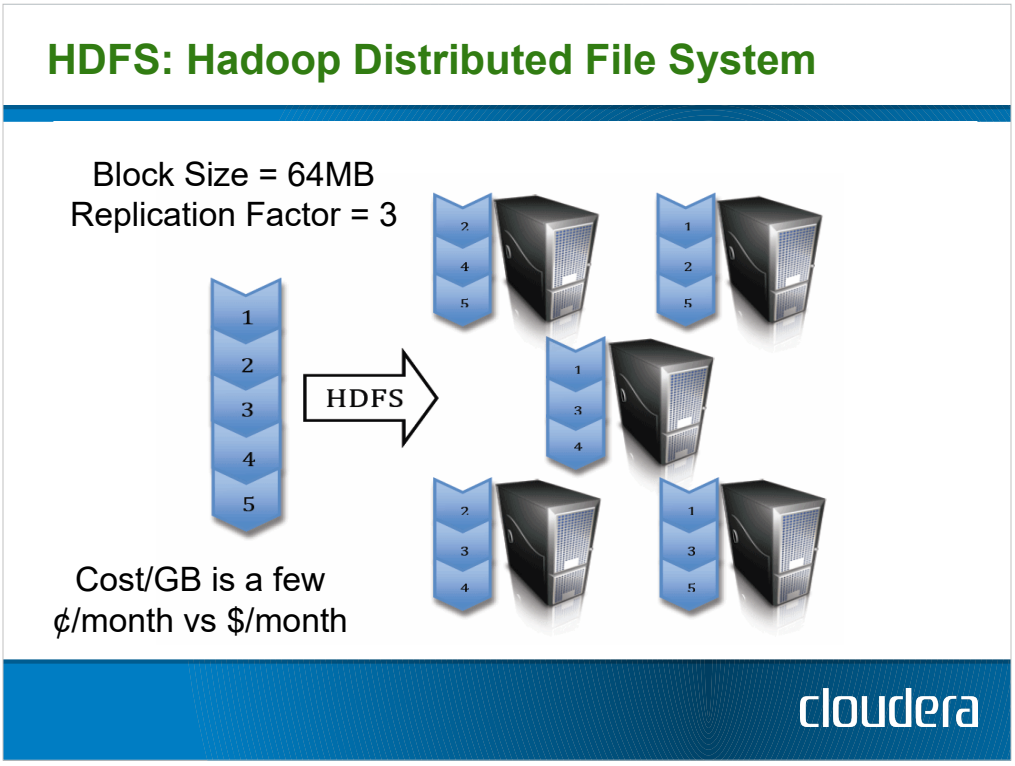
- Networks industry standard hardware nodes together to combine compute and storage into scalable distributed system
- Scales to petabytes without modification
- Manages fault tolerance and data replication automatically
- Processes semi-structured and unstructured data easily
- Supports MapReduce natively to analyze data in parallel

cloudera

What Hadoop does not do

- No random access or transactions
  - **Hadoop is not a database**
- Not real-time
  - **Hadoop is batch oriented**
- Push-button install or configure
  - **Hadoop requires some advanced skills**
  - This aspect is part of Cloudera's value add

cloudera



- Pool commodity servers in a single hierarchical namespace.
- Designed for large files that are written once and read many times.
- Example here shows what happens with a replication factor of 3, each data block is present in at least 3 separate data nodes.
- Typical Hadoop node is eight cores with 16GB ram and four 1TB SATA disks.
- Default block size is 64MB, though most folks now set it to 128MB

MapReduce

- In simple terms, it's an application with 2 functions
  - Map Function
    - Think massively parallel "**group by**"
  - Reduce Function
    - Think "**aggregation + processing**"
- Not hard to write
  - Can be challenging to refactor existing algorithms
- Designed to work hand-in-hand with HDFS
  - Minimizes disk seeks, operates at "transfer rate" of disk

cloudera

Note: these are not simple queries, they are DEEP COMPLEX SCANS

Speed @ Scale

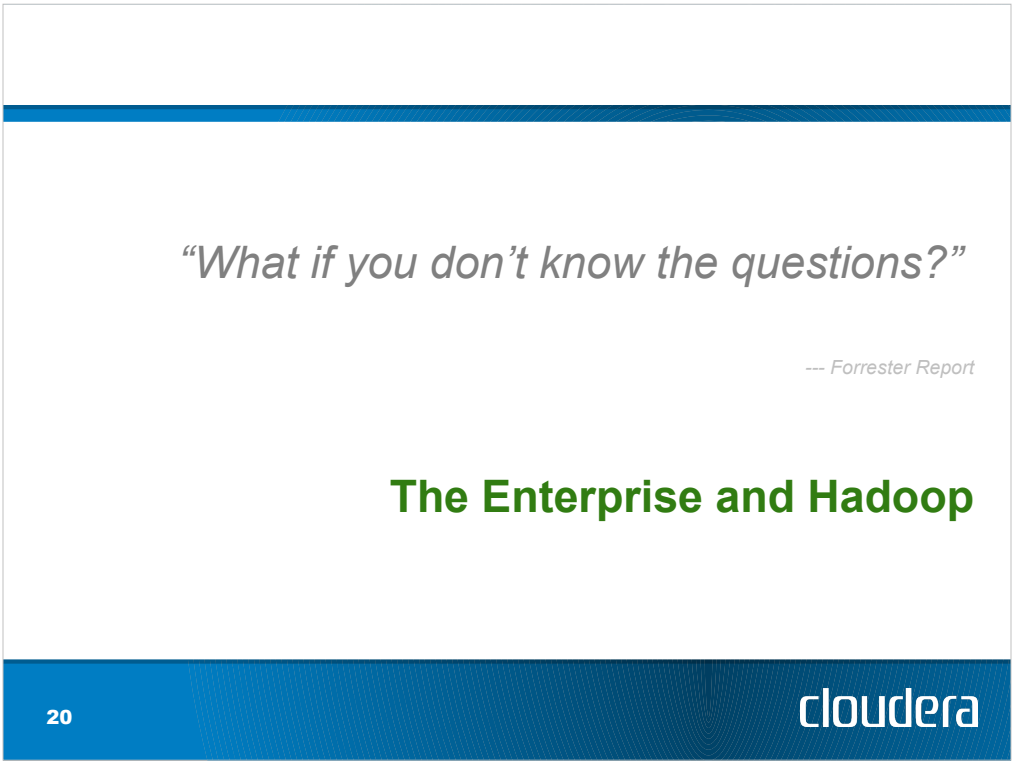
- Scenario
  - 1 million sensors, collecting sample / 5 min
  - 5 year retention policy
  - Storage needs of 15 TB
- Processing
  - Single Machine: 15TB takes 2.2 DAYS to scan
  - MapReduce@ 20 nodes: **Same task takes 11 Minutes**

cloudera

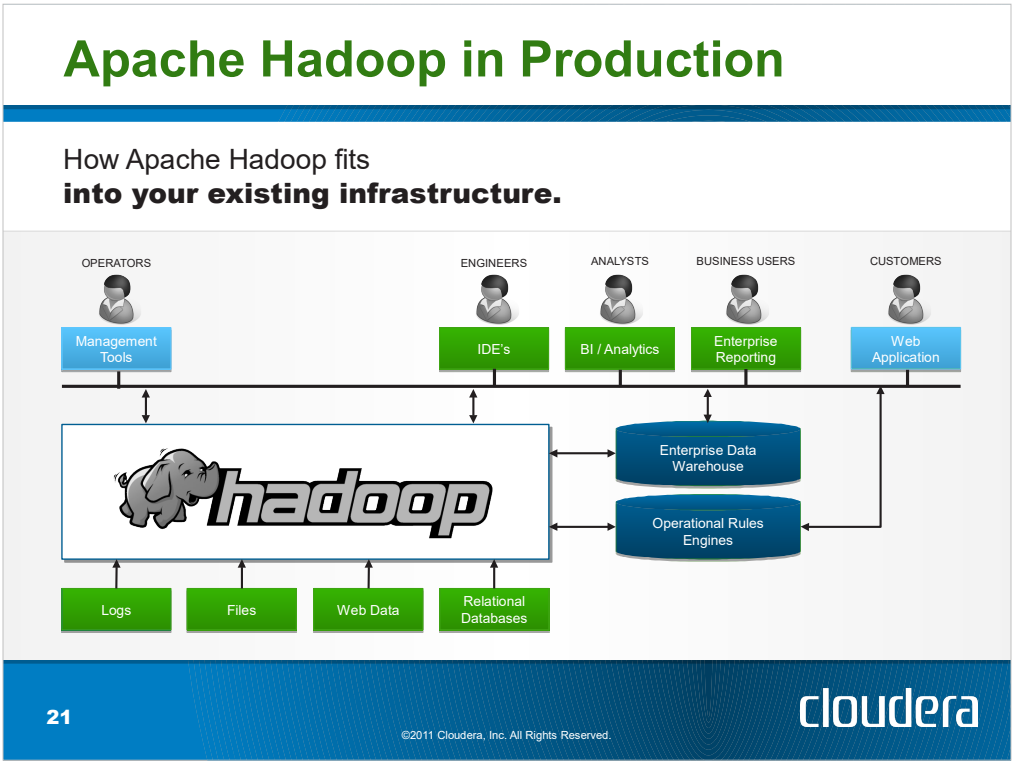
Note: these are not simple queries, they are DEEP COMPLEX SCANS

MapReduce Tools

- Pig
  - Procedural language compiled into MR
- Hive
  - **SQL-like** language compiled into MR
- Mahout
  - Collection of data mining algorithms for Hadoop
- Streaming
  - Ability to write MR with tools such as python, etc



Theme: they through away a lot of valuable gas and oil just like we through away data today



Apache Hadoop is a new solution in your existing infrastructure.

It does not replace any existing major existing investment.

Apache brings data that you're already generating into context and integrates it with your business.

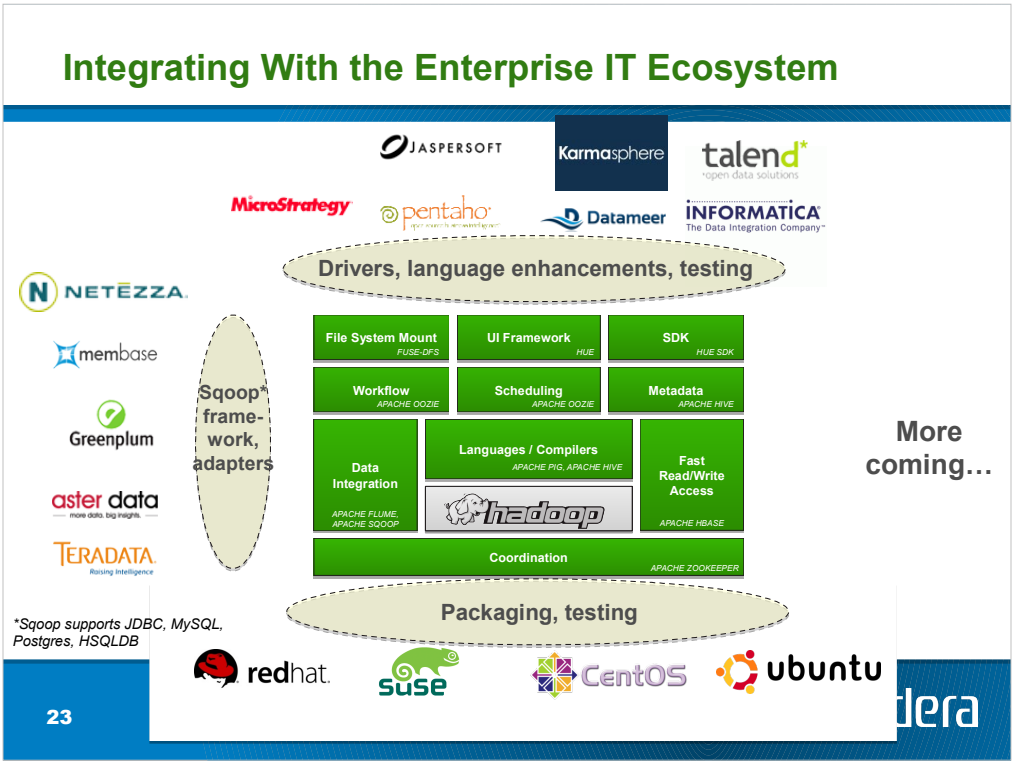
You get access to key information about how your business is operating but pulling together

- Web and application logs
- Unstructured files
- Web data
- Relational data

Hadoop is used by your team to analyze this data and deliver it to business users directly and via existing data management technologies

What the Industry is Doing

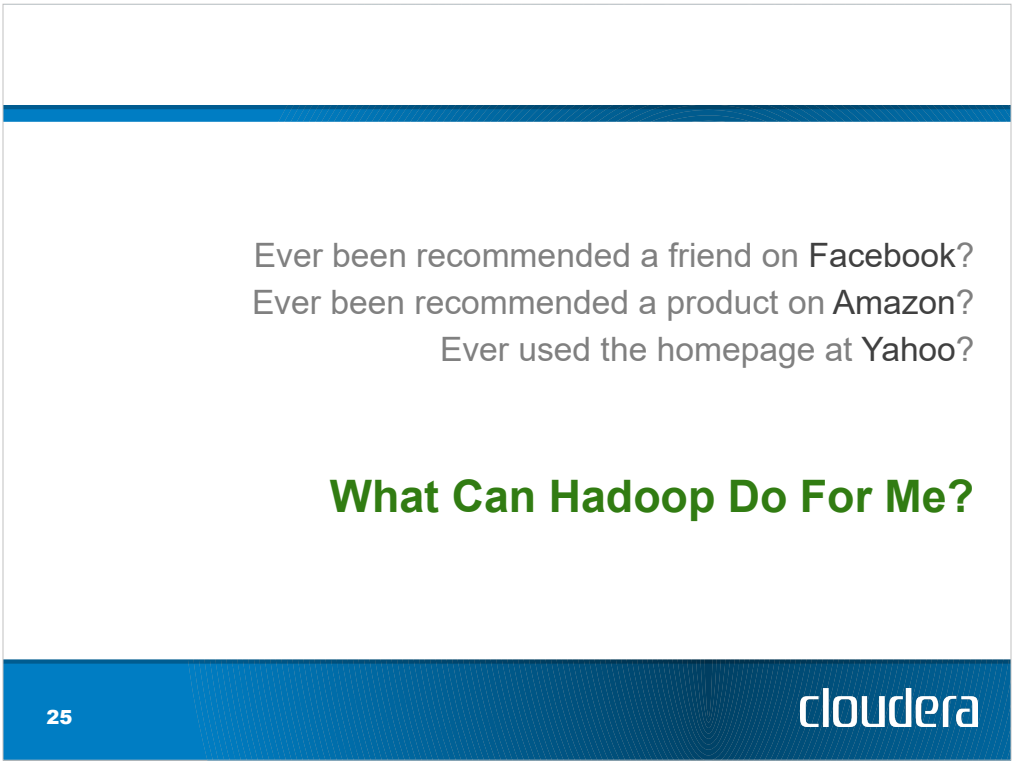
- Microsoft
  - ships CTP of **Hadoop Connectors for SQL Server and Parallel Data Warehouse**
    - (based on Cloudera's Sqoop)
  - http://blogs.technet.com/b/dataplatforminsider/archive/2011/08/25/microsoft-ships-ctp-of-hadoop-connectors-for-sql-server-and-parallel-data-warehouse.aspx
- Oracle
  - Announcments at Oracle Open World
    - Connector to **Hadoop** allowing data flow into Oracle
    - **Hadoop** Accelerator for Exalogic
    - In-memory processing for MapReduce
    - ETL using **Hadoop**
    - Integrated analytics on Oracle and **Hadoop**



1. Allows vendor specific solutions to build on top
2. Certain custom code would be written for the platform as well

Forester Report

- World Economic Forum
  - **Declared that data is a new asset class**
- Big data is an **applied science project** in most companies
  - Major potential constraint is **not** the cost of the computing technology
  - **but** the **skilled people** needed to carry out these projects
    - the data scientists
- What if you don't know the questions?
  - Big data is all about exploration without preconceived notions
  - Need tools to ask questions to understand the right questions to ask
- Much of the software is based on open-source **Hadoop**
  - http://blogs.forrester.com/brian_hopkins/11-09-30-big_data_will_help_shape_your_markets_next_big_winners



Theme: they through away a lot of valuable gas and oil just like we through away data today

Problems Addressed With Hadoop

- Text Mining
- Index Building
  - Search Engines
- Graph Creation
  - Twitter, Facebook
- Pattern Recognition
  - Naïve Bayes Classification
- Recommendation Engines
- Predictive Models
- Risk Assessment

cloudera

Transition into “hadoop as a platform”, but vendors are building on top of it for specific vertical challenges

A Few Named Examples

Analyze search terms and subsequent user purchase decisions to tune search results, increase conversion rates

Digest long-term historical trade data to identify fraudulent activity and build real-time fraud prevention

Model site visitor behavior with analytics that deliver better recommendations for new purchases

Continually refine predictive models for advertising response rates to deliver more precisely targeted advertisements

Replace expensive legacy ETL system with more flexible, cheaper infrastructure that is 20 times faster

Correlate educational outcomes with programs and student histories to improve results

cloudera

Packages For Hadoop

- DataFu
  - From Linkedin
    - http://sna-projects.com/datafu/
  - UDFs in Pig
  - used at LinkedIn in many of off-line workflows for data derived products
    - "People You May Know"
    - "Skills"
  - Techniques
    - PageRank
    - Quantiles (median), variance, etc.
    - Sessionization
    - Convenience bag functions
    - Convenience utility functions

Integration with Libs

- Mix MapReduce with Machine Learning Libs
  - WEKA
  - KXEN
  - CPLEX
- Map side “groups data”
- Reduce side processes groups of data with Lib in parallel
  - Involves tricks in getting K/V pairs into lib
  - Pipes, tmp files, task cache dir, etc

Ask the right questions up front..

- Is the job disk bound?
- What is the latency requirement on the job?
  - Does it need a sub-minute latency? (not good for Hadoop!)
- Does the job look at a lot or all of the data at the same time?
  - Hadoop is good at looking at all data, complex/fuzzy joins
- Is large amounts of ETL processing needed before analysis?
  - Hadoop is good at ETL pre-processing work
- Can the analysis be converted into MR / Pig / Hive?

cloudera

This is probably a better slide 30

What Hadoop Not Good At in Data Mining

- Anything highly iterative
- Anything that is extemely CPU bound and not disk bound
- Algorithms that can't be inherently parallelized
  - Examples
    - Stochastic Gradient Descent (SGD)
    - Support Vector Machines (SVM)
  - Doesn't mean they arent great to use

cloudera

Check this against the Mahout impl

Questions? (Thank You!)

- **Hadoop World 2011**
  - http://www.hadoopworld.com/
- Cloudera's Distribution including Apache Hadoop (CDH):
  - http://www.cloudera.com
- Resources
  - http://www.slideshare.net/cloudera/hadoop-as-the-platform-for-the-smartgrid-at-tva
  - http://gigaom.com/cleantech/the-google-android-of-the-smart-grid-openpdc/
  - http://news.cnet.com/8301-13846_3-10393259-62.html
  - http://gigaom.com/cleantech/how-to-use-open-source-hadoop-for-the-smart-grid/
- Timeseries blog article
  - http://www.cloudera.com/blog/2011/03/simple-moving-average-secondary-sort-and-mapreduce-part-1/

More?

- Look at www.cloudera.com/training to learn more about Hadoop
- Read www.cloudera.com/blog
  - Lots of great use cases.
- Check out the downloads page at
  - www.cloudera.com/downloads
  - Get your own copy of Cloudera Distribution for Apache Hadoop (CDH)
  - Grab Demo VMs, Connectors, other useful tools.
- Contact Josh with any questions at
  - **josh@cloudera.com**