Getting the Big (Data) Picture

Eva Andreasson , Cloudera

Big Data?

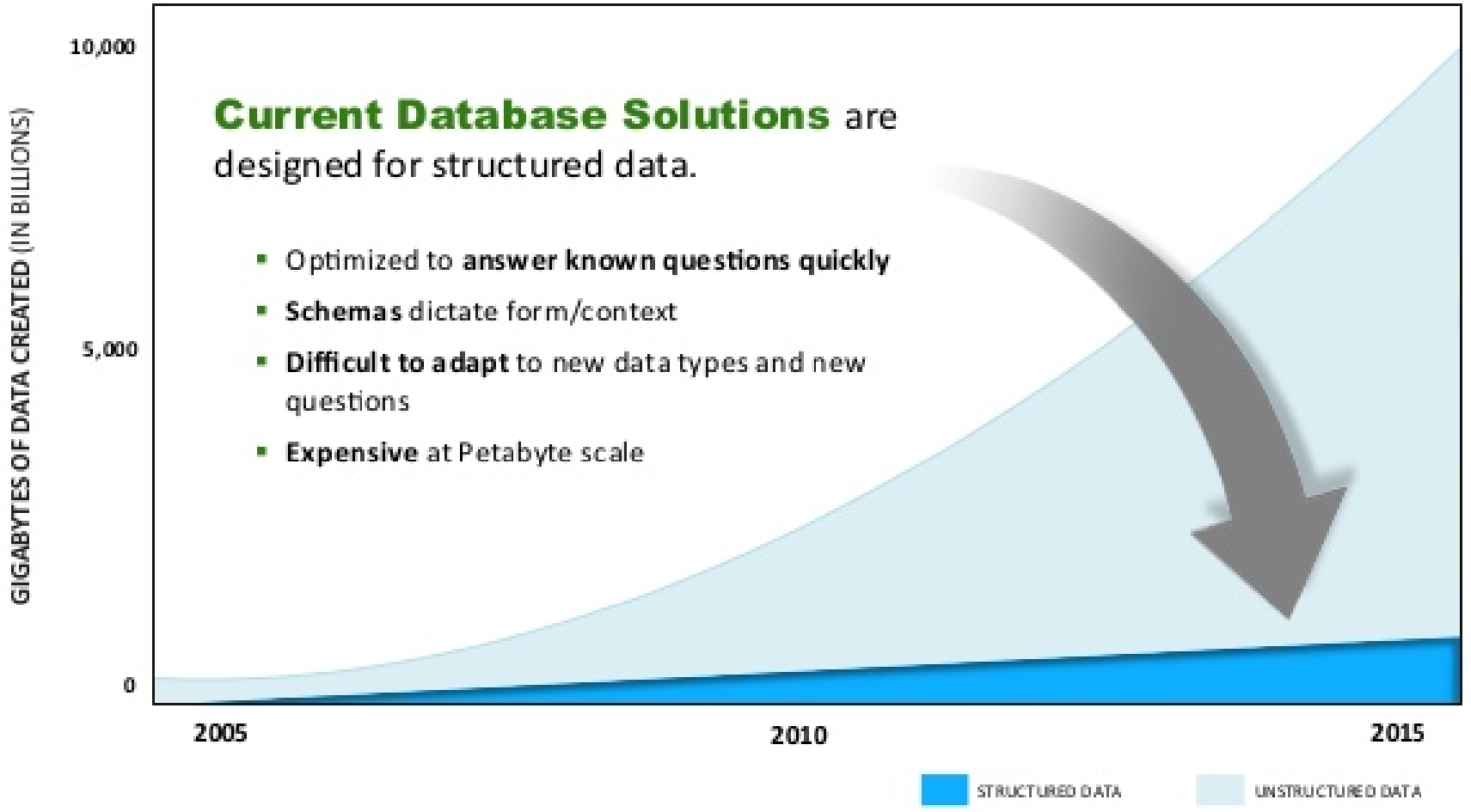
Today's Big Data Landscape Journey

- PART 1 – 10000ft
  - Drivers to re-thinking data
  - Where does Hadoop come from?
  - Industry trends and vendor map
  - When should I use which tool?
- PART 2 – Back to Earth
  - Walk through of a big data use case
- Q&A
- Break
- PART 3 – Deep Dive
  - Dean Wampler deep diving on Spark and the comeback of SQL

Big Data Evolution

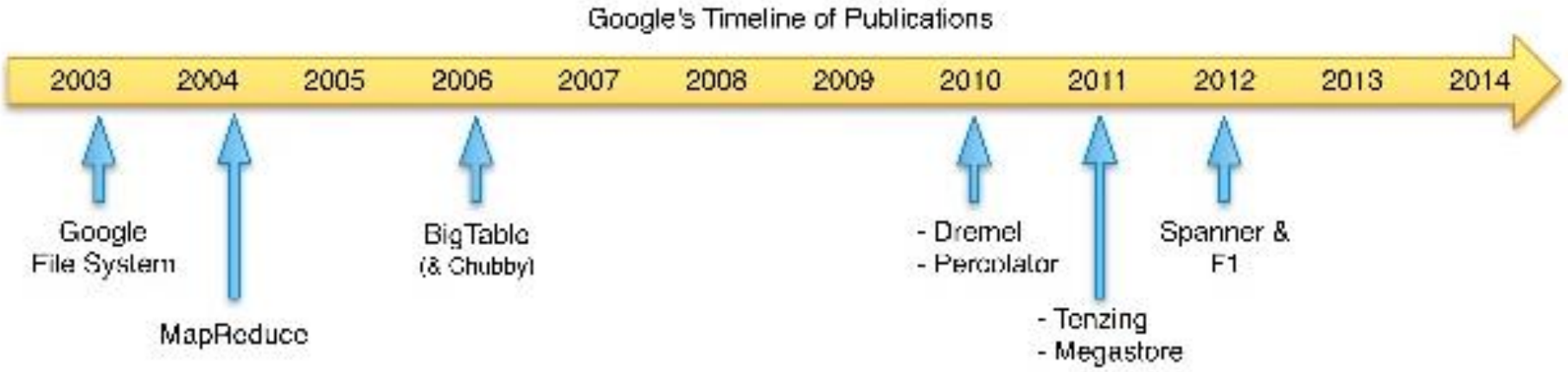
Data Re-Thinking Drivers

Existing Technology Failing?

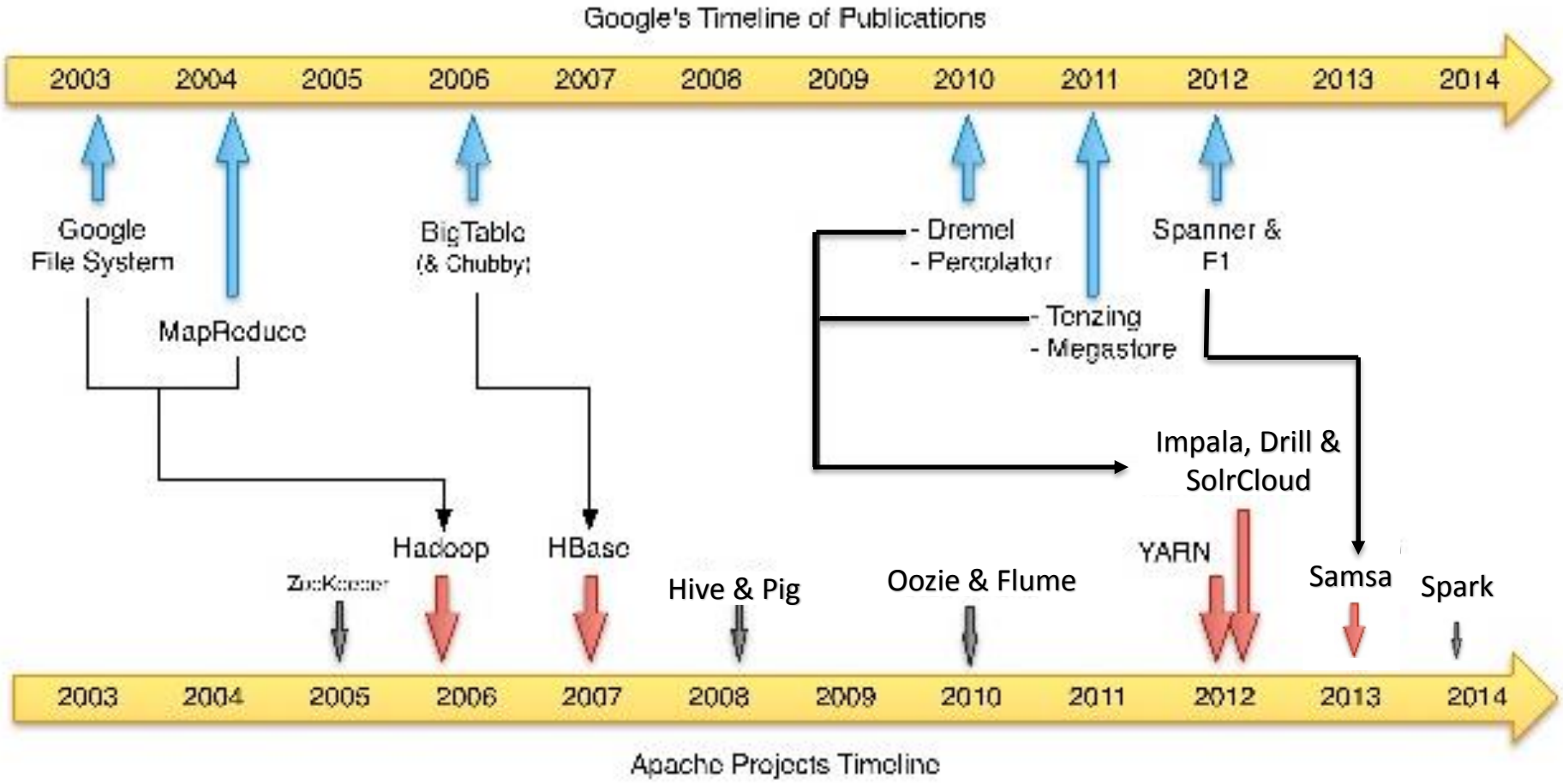


“A smart engineer comes up with great a solution. A wise engineer knows to ‘Google’ it first…”

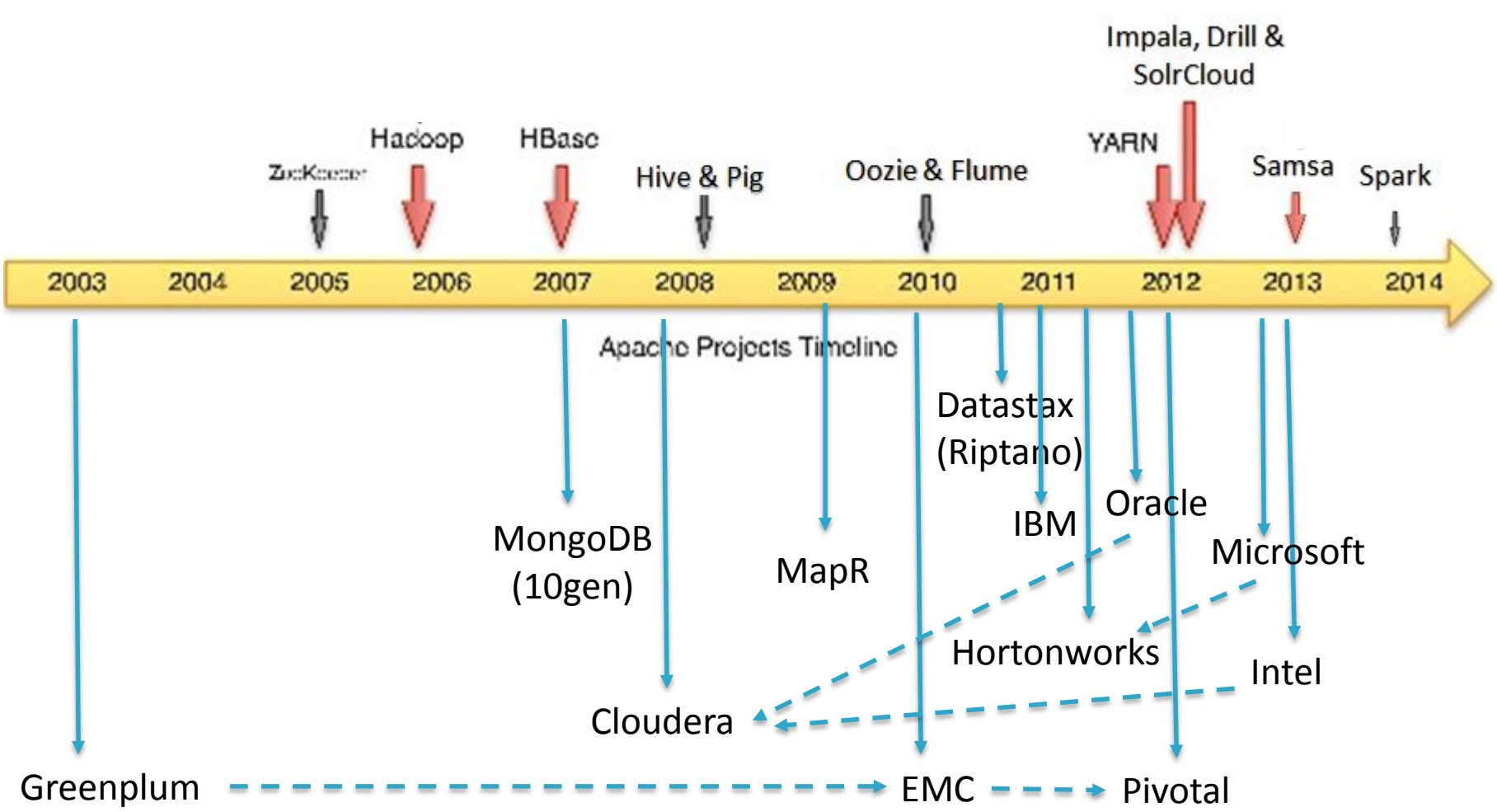
Technology Evolution



Technology Evolution



Hadoop Distribution Vendor Evolution



Snapshot of the Data Management Landscape
(NOTE: Borders are Fuzzy, Not Exhaustive Lists)

APPLICATION

BI / Visualization / Analytics Tools

- 0xData
- Alteryx
- AVATA
- Datameer
- IBM
- Karmasphere
- Opera
- Oracle
- Palantir
- Platfora
- SAP
- SAS
- Tableau
- Tibco
- Trifacta
- Microsoft
- Microstrategy
- Qlickview
- Teradata Aster
- Zoomdata

INFRASTRUCTURE

Analytics

- *Cloudera*
- *Hadapt*
- *Hortonworks*
- *Infobright*
- *Kognito*
- *MapR*
- *Netezza*
- *Pivotal*

Operational

- *Couchbase*
- *Datastax*
- *Informatica*
- *MarkLogic*
- *MongoDB*
- *Splunk*
- *Terracotta*
- *VoltDB*

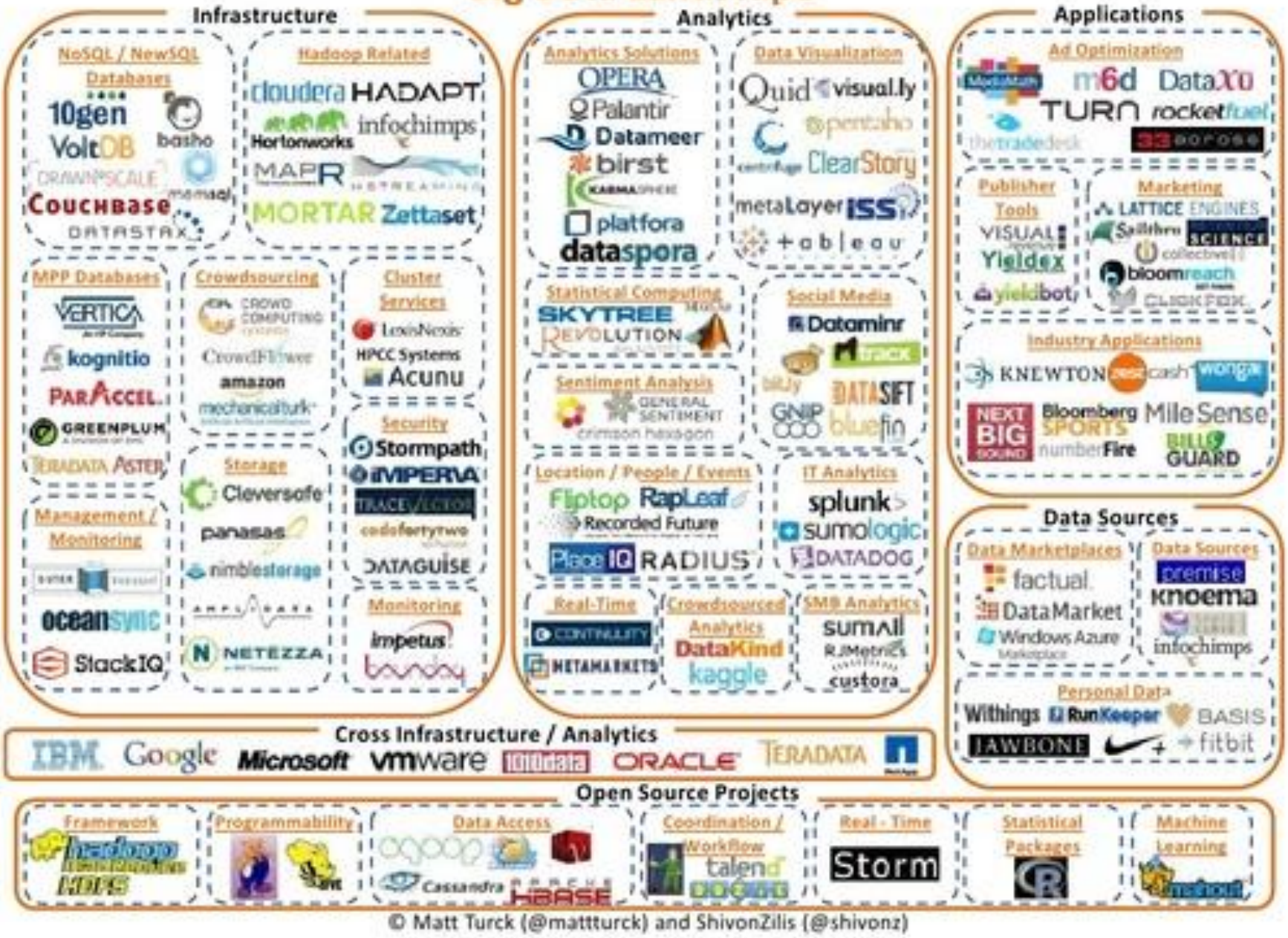
As A Service

- *Amazon web services*
- *CSC*
- *Google BigQuery*
- *Mortar*
- *Quobole*
- *Windows Azure*

Structured DB

- *IBM DB2*
- *MemSQL*
- *MySQL*
- *Oracle*
- *PostgreSQL*
- *SQLServer*
- *Sybase*
- *Terradata*

Open Source Technology



© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

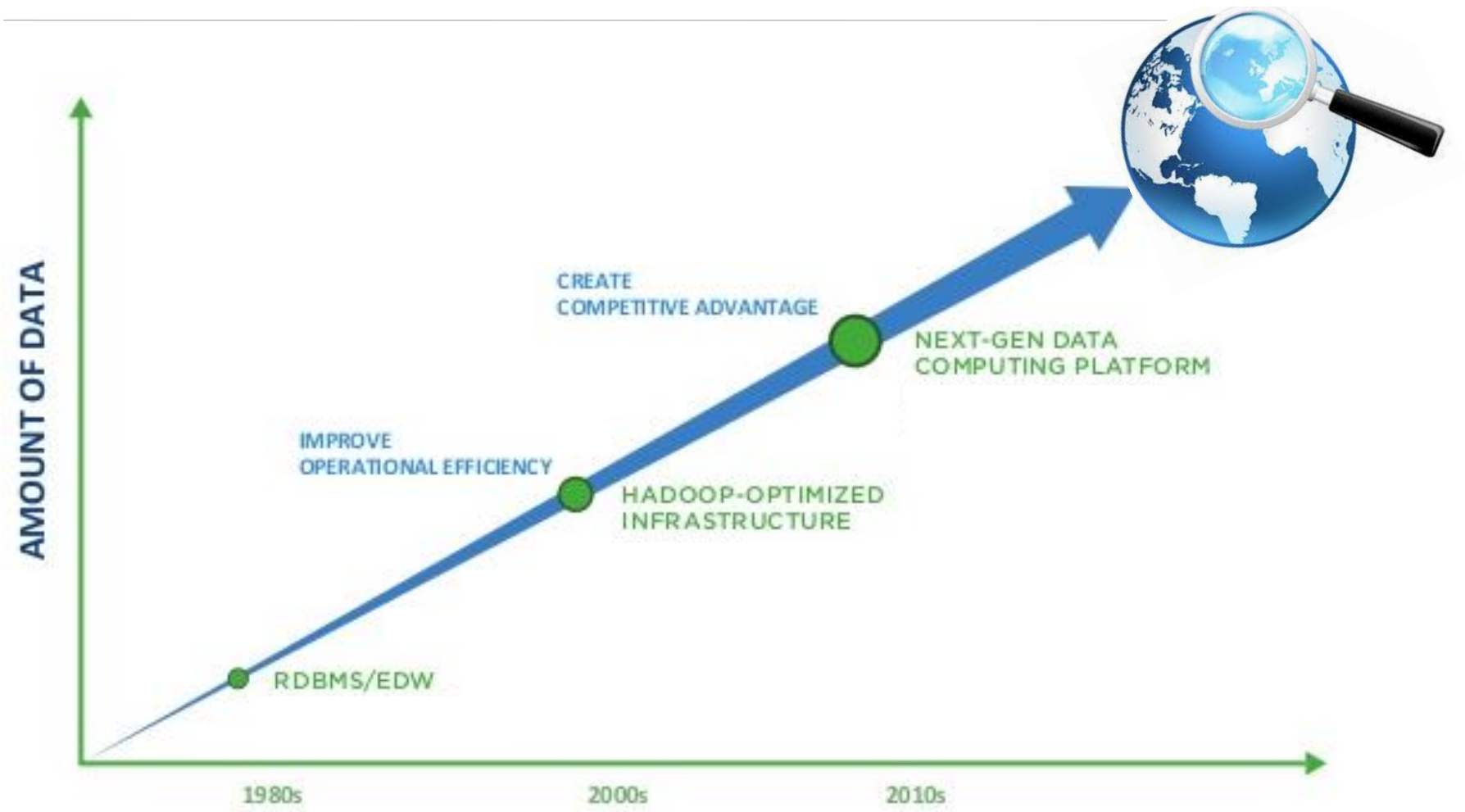
It is Here to Stay…

2013

2014

New Organizational Data Needs also Drive IT Architecture Evolution

Where we are Heading…

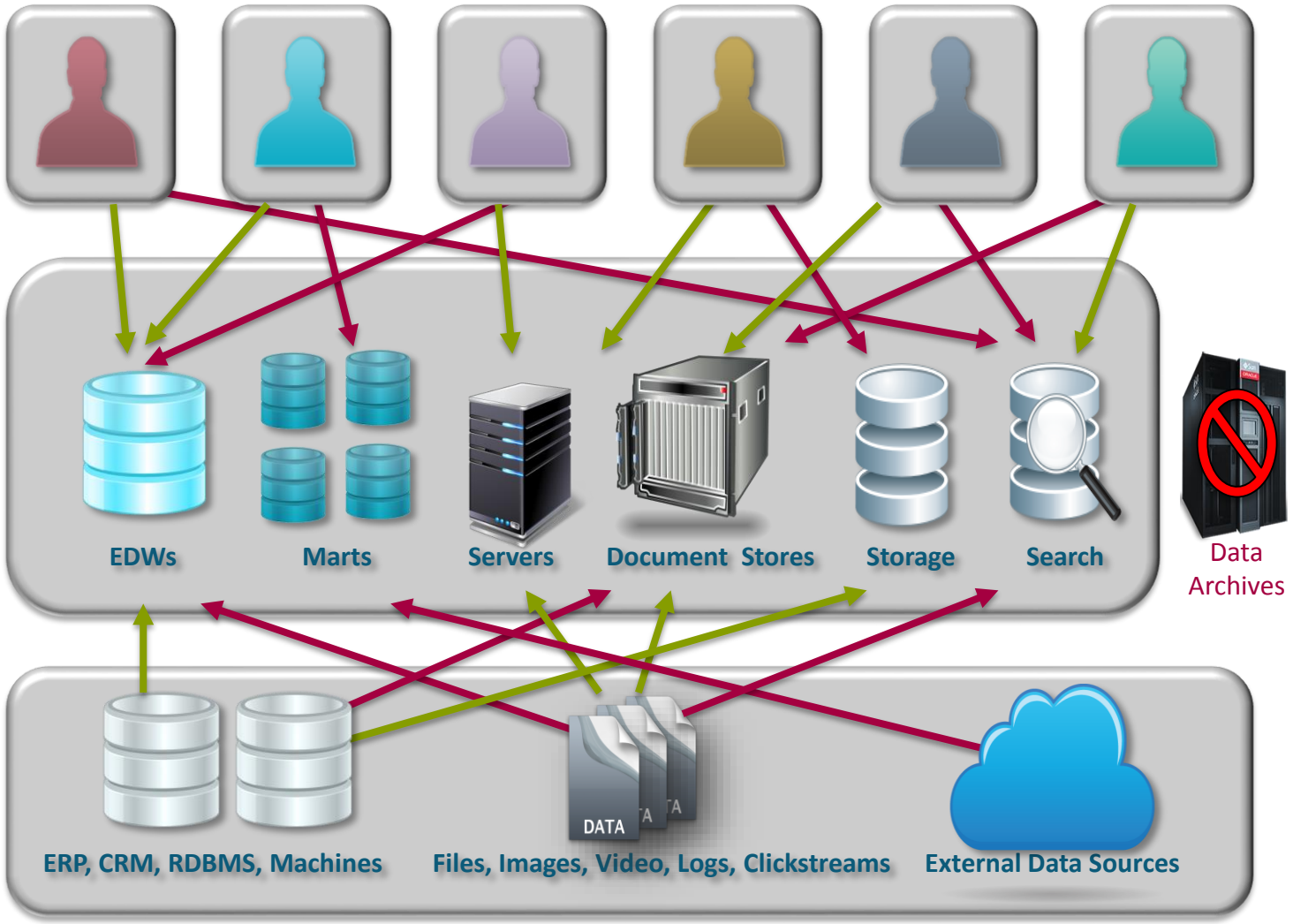


The Need to Rethink Data Architecture

Thousands of Employees & Lots of Inaccessible Information

Heterogeneous Legacy IT Infrastructure

Silos of Multi-Structured Data Difficult to Integrate

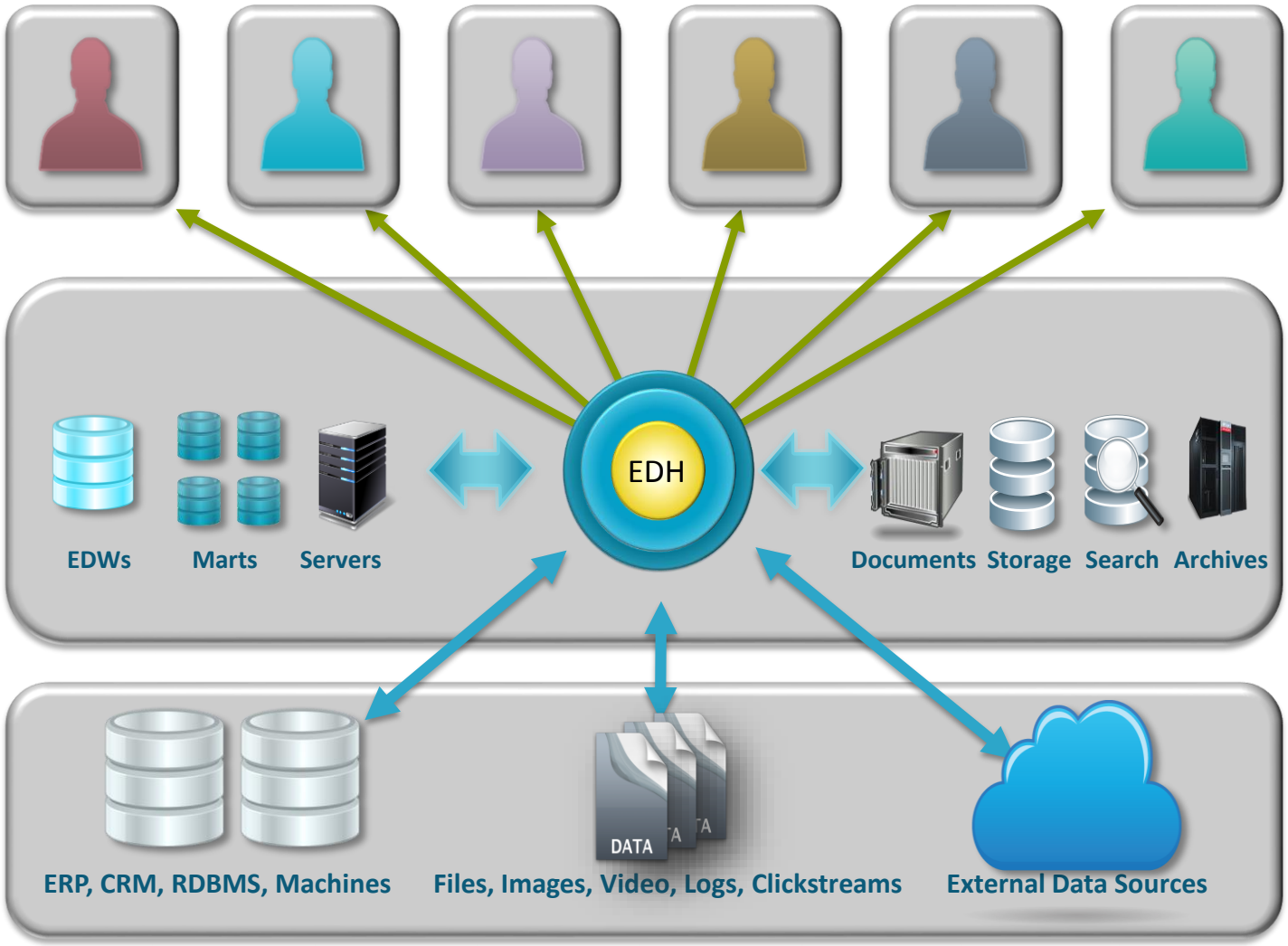


New Category: The Enterprise Data Hub (EDH)

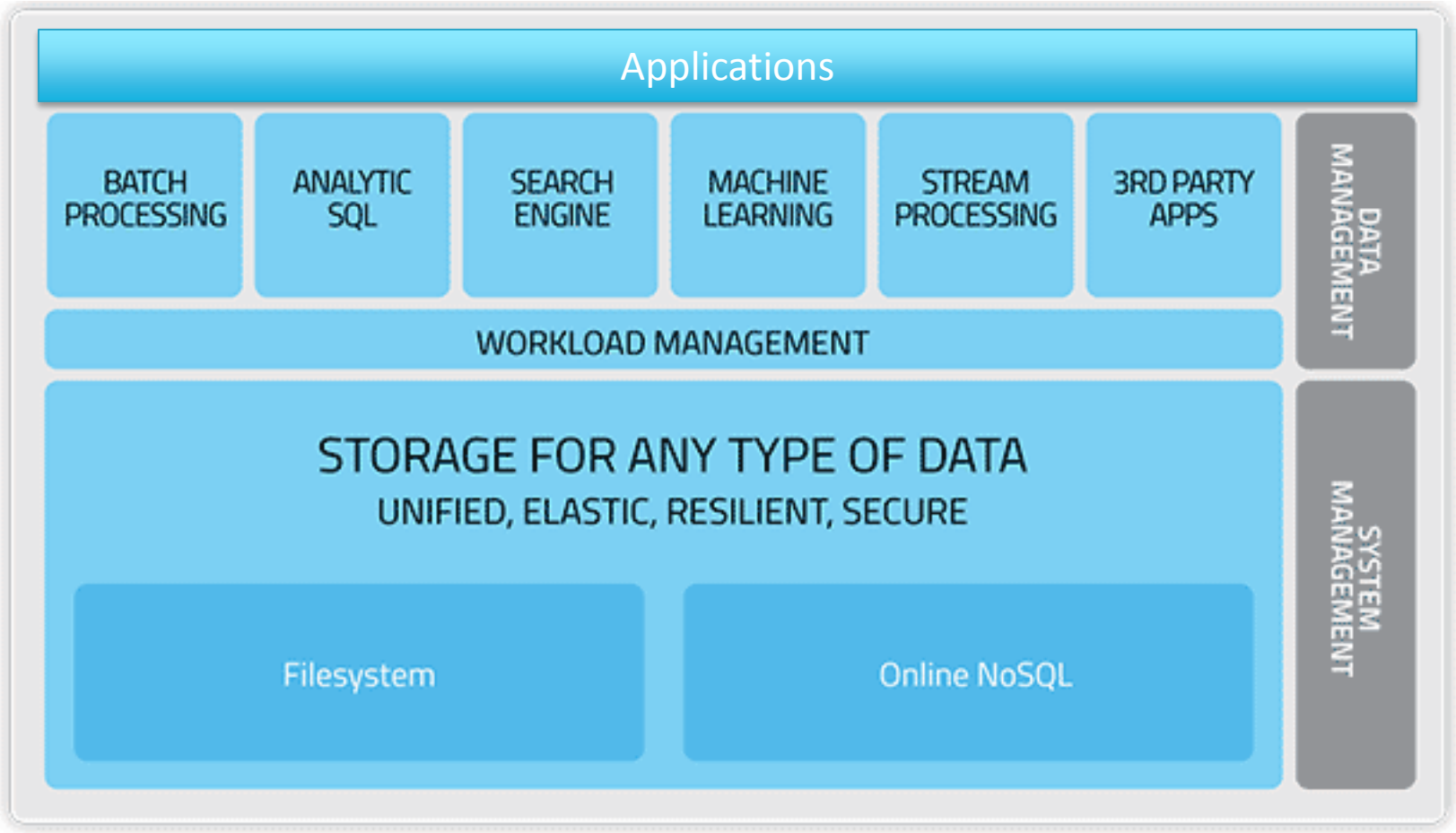
Information & data accessible by all for insight using leading tools and apps

Enterprise Data Hub
Unified Data Management Infrastructure

Ingest All Data
Any Type
Any Scale
From Any Source



Hadoop et al Enabling an EDH



The Right Tool for the Right Task

When to use what?

- Real Time Query (e.g. Impala)
  - I want to do BI reports or interactive analytical aggregations but not wait hours for the response
- Batch Query (e.g. Pig, Hive)
  - I have nightly batch query jobs as part of a workflow
- Real Time Search (e.g. SolrCloud)
  - I have unstructured data I want to free text over
  - My SQL queries are getting more and more complex as they need to contain 15+ “like” conditions
- Real time key lookups (e.g. Hbase)
  - I want random access to sparsely populated table-like data
  - I want to compare user profiles or behavior in real time

When to use what?

- Spark
  - I want to implement analytics algorithms over my data, and my data sets fit into memory
  - I have real time streaming data I want to analyze in real time
- MapReduce
  - I want to do fail-safe large ETL processing workloads
  - My data does not fit into memory and I want to batch process it with my custom logic – no real time needs

PART 2: Let’s Make it Real

Introducing “DataCo”

- A product and service provider
- Medium sized
- Most revenue via online store
- Customer transactions stored in an RDBMS
- Business as usual, but market is getting more competitive

- Pretty much any company?

"I only have ~100GB. I don't have a Big Data problem." – Head of IT, DataCo

Now…

- Pretend you work for the Head of IT
- Pretend you are pretty smart… ☺
- Assume you have a 10 node CDH cluster running (in AWS?) just for fun..
  - CDH = Clousera's Distribution incl. Apache Hadoop

BQ1: What products should we invest in?

- First step:
  - Try something you already know how to do
  - Do the *same* product sales report, but in CDH
- Approach:
  - Load product sales data into HDFS from RDBMS, using Sqoop
  - Convert data to Avro (to optimize for any future workload)
  - Create Hive tables to serve the question at hand
  - Use Impala to query (you don't want to wait forever…)
  - Find out the top 10 most sold products

Same use cases in a platform that scales with data growth

Example Sqoop Ingest Job from MySQL

- Log into your Master Node via SSH and Sqoop in data

```
$ sqoop import-all-tables  -m 12 –connect
jdbc:mysql://my.sql.host:3306/retail_db  --username=dataco_dba
--password=goto2014  --compression-codec=snappy  --as-avrodatafile
--warehouse-dir=/user/hive/warehouse
```

- View your imported tables

```
$ hadoop fs -ls /user/hive/warehouse/
```

- View all Avro files constituting the “Categories” table

```
$ hadoop fs -ls /user/hive/warehouse/categories/
```

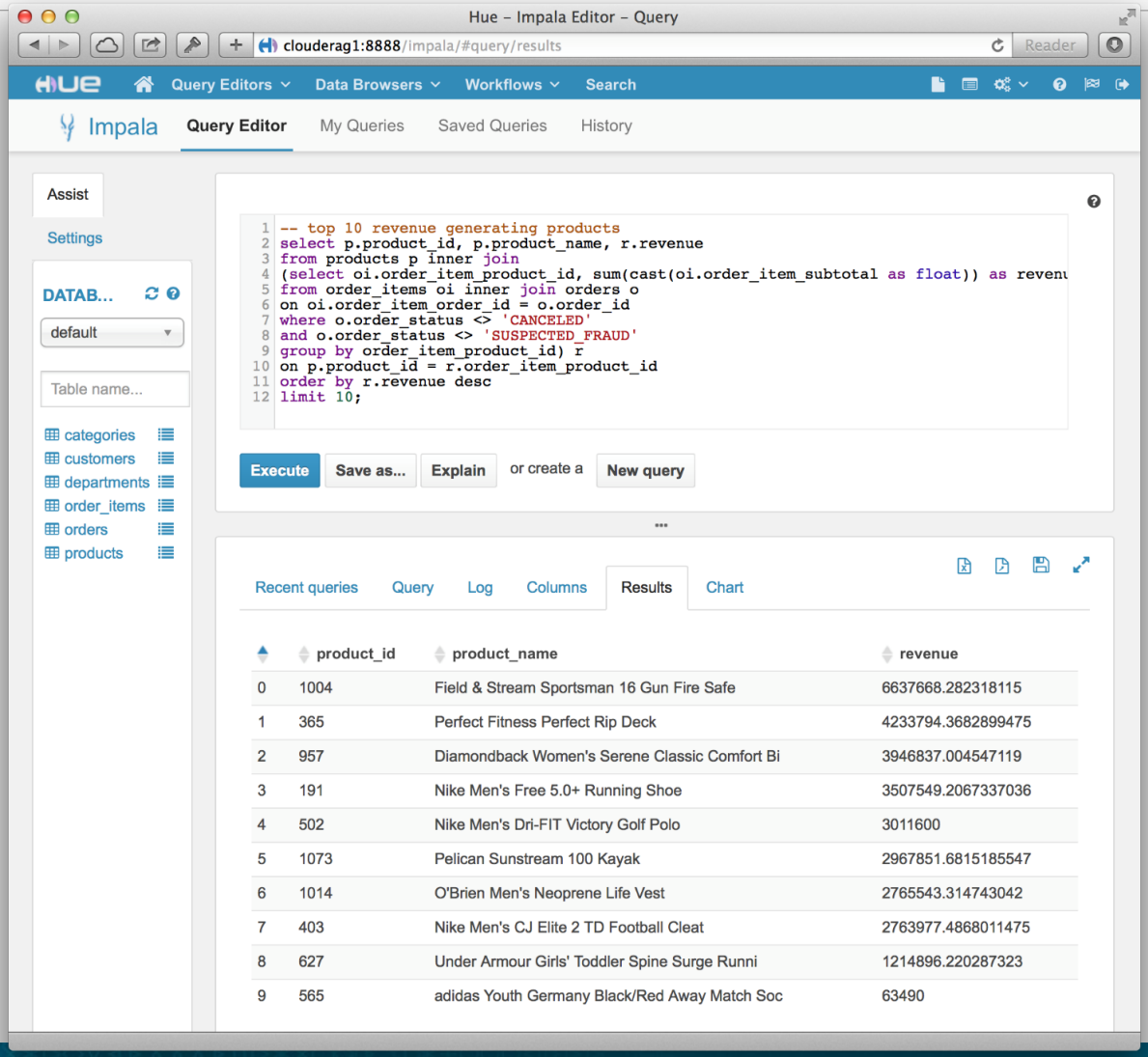
Create Tables in Hive

- Create tables in Hive to serve the query at hand

```
hive> CREATE EXTERNAL TABLE products
> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.avro.AvroSerDe'
> STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat'
> OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat'
> LOCATION 'hdfs:///user/hive/warehouse/products'
> TBLPROPERTIES
('avro.schema.url'='hdfs://namenode_dataco/user/examples/products.avsc');
```

- NOTE: You will need more tables than the example above to serve the query…

Use Impala via Hue to Query



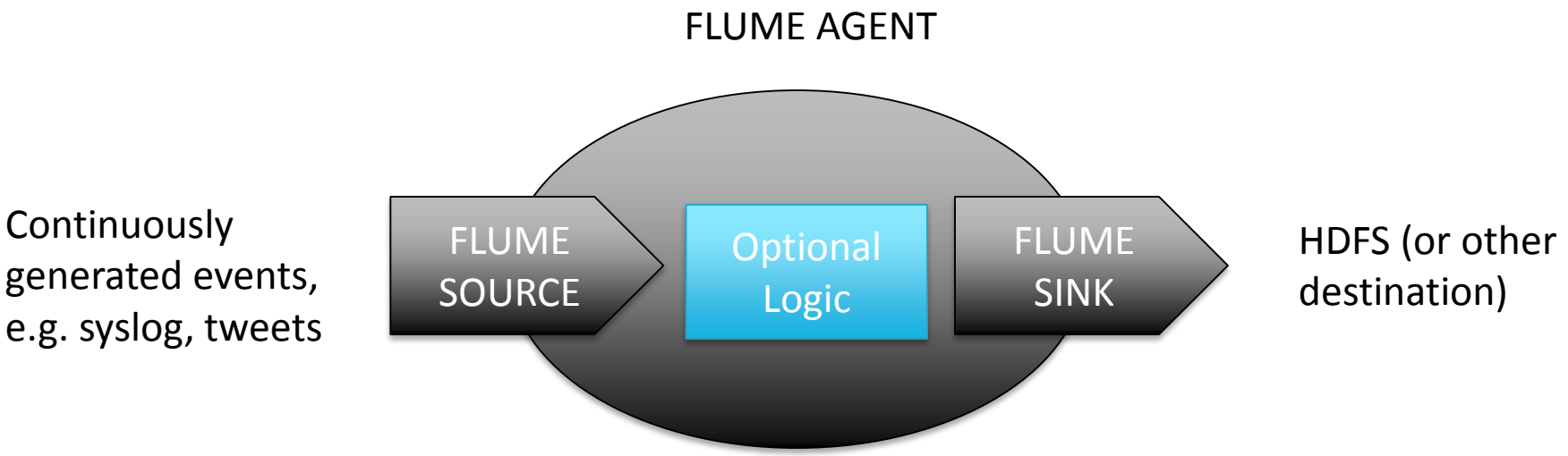
BQ1: What products should we invest in?

- Second step:
  - Get “big data” value by analyzing multiple data sets to serve the same business question
- Approach:
  - Load web log data into the same platform
  - Create Hive tables over semi-structured view events
  - Use Hue and Impala to query
  - Find out the top 10 most *viewed* products

Multiple data sets give better insight = Big Data value

Ingest Data Using Flume

- Pub/sub ingest framework
- Flexible multi-level (mini-transformation) pipeline



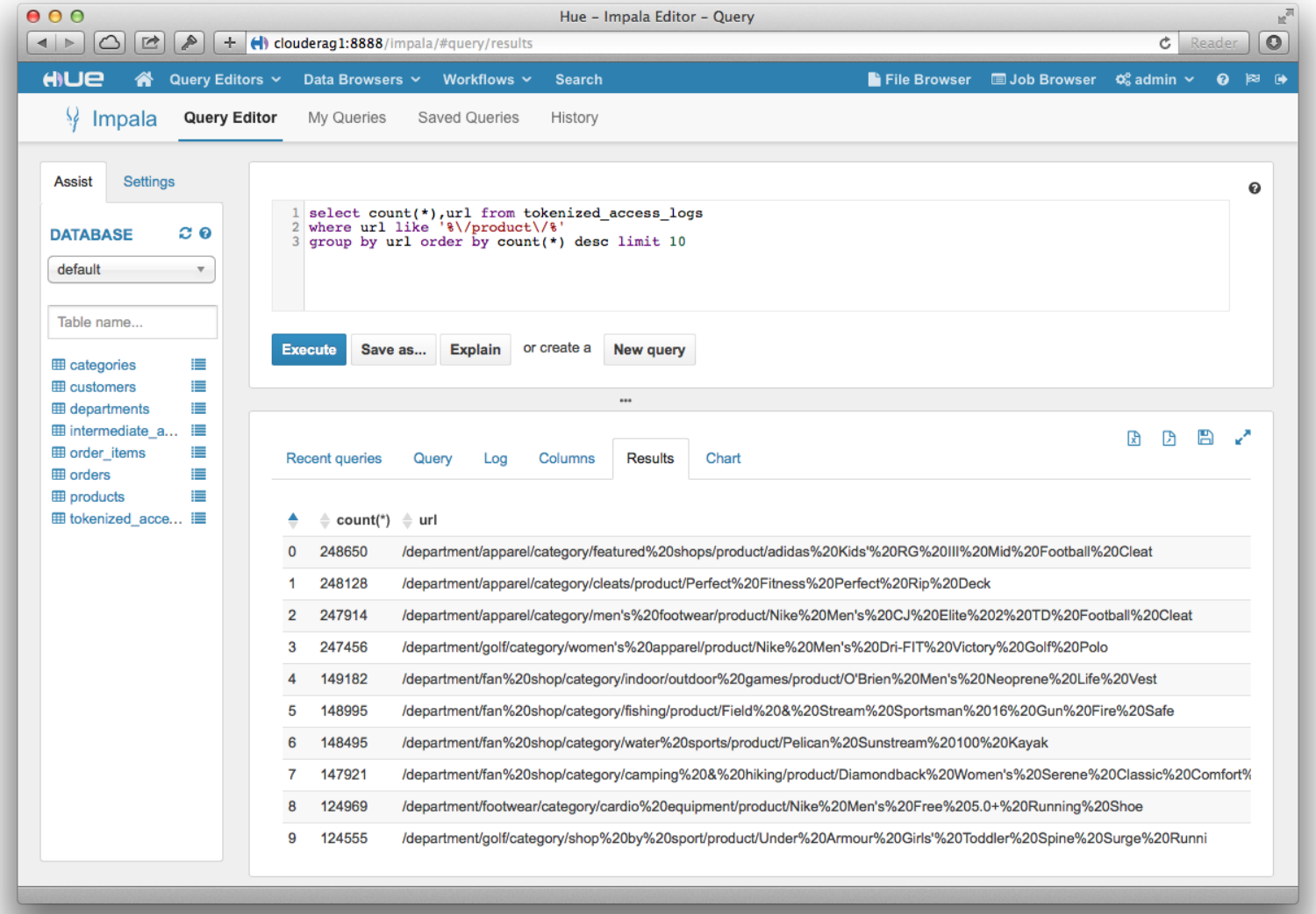
Create Hive Tables over Log Data

- Ingest data using Flume
- Create new tables over log data to serve the same BQ

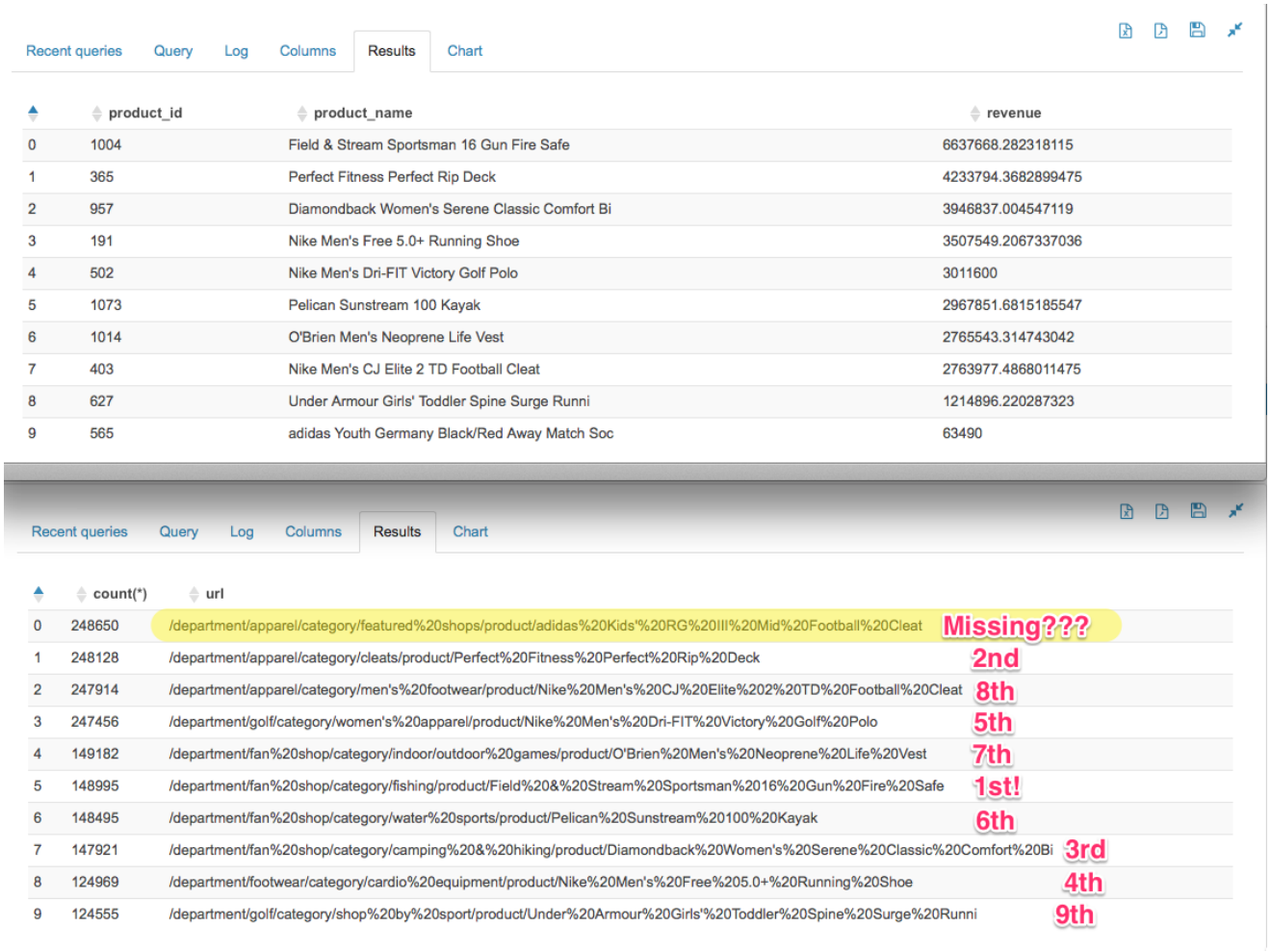
```
CREATE EXTERNAL TABLE intermediate_access_logs ( ip STRING, date STRING,
method STRING, url STRING, http_version STRING, code1 STRING, code2 STRING,
dash STRING, user_agent STRING) ROW FORMAT SERDE
'org.apache.hadoop.hive.contrib.serde2.RegexSerDe' WITH SERDEPROPERTIES (
"input.regex" = "([^ ]*) - - \\[([^\\]]*)\\] \"([^\ ]*) ([^\ ]*) ([^\ ]*)\" (\\d*) (\\d*) \"([^\"]*)\" \"([^\"]*)\"",
"output.format.string" = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s" )
LOCATION '/user/hive/warehouse/original_access_logs';

CREATE EXTERNAL TABLE tokenized_access_logs ( ip STRING, date STRING, method
STRING, url STRING, http_version STRING, code1 STRING, code2 STRING, dash
STRING, user_agent STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/user/hive/warehouse/tokenized_access_logs'; ADD JAR
/opt/cloudera/parcels/CDH/lib/hive/lib/hive-contrib.jar; INSERT OVERWRITE TABLE
tokenized_access_logs SELECT * FROM intermediate_access_logs; exit;
```

Use Impala and Hue to Query



Most Viewed List Differ from Most Sold???



Recent queries | Query | Log | Columns | **Results** | Chart

	product_id	product_name	revenue
0	1004	Field & Stream Sportsman 16 Gun Fire Safe	6637668.282318115
1	365	Perfect Fitness Perfect Rip Deck	4233794.3682899475
2	957	Diamondback Women's Serene Classic Comfort Bi	3946837.004547119
3	191	Nike Men's Free 5.0+ Running Shoe	3507549.2067337036
4	502	Nike Men's Dri-FIT Victory Golf Polo	3011600
5	1073	Pelican Sunstream 100 Kayak	2967851.6815185547
6	1014	O'Brien Men's Neoprene Life Vest	2765543.314743042
7	403	Nike Men's CJ Elite 2 TD Football Cleat	2763977.4868011475
8	627	Under Armour Girls' Toddler Spine Surge Runni	1214896.220287323
9	565	adidas Youth Germany Black/Red Away Match Soc	63490

Recent queries | Query | Log | Columns | **Results** | Chart

	count(*)	url	
0	248650	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat	Missing???
1	248128	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck	2nd
2	247914	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat	8th
3	247456	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo	5th
4	149182	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest	7th
5	148995	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe	1st!
6	148495	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak	6th
7	147921	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi	3rd
8	124969	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe	4th
9	124555	/department/golf/category/shop%20by%20sport/product/Under%20Armour%20Girls'%20Toddler%20Spine%20Surge%20Runni	9th

BQ2: Why is sales suddenly dropping?

- Third Step
  - Use same data to serve multiple use cases
  - EDH value: multiple business needs in the same platform, without moving data
- Approach
  - Use same web log data
  - Index it at ingest using Flume and SolrCloud
    - Create a Solr collection and an index schema
    - Configure the Flume agent to parse incoming data into the index schema, using Morphlines
  - Search via Hue and resolve issues over real-time data

Multiple use cases over same data without data move = EDH value

Create your Index

- Create an empty Solr index configuration directory

```
$ solrctl --zk <ALL YOUR ZK IPs>/solr instancedir --generate live_logs_dir
```

- Edit the Solr Schema file to have the fields you want to search over

```
…
<field name="_version_" type="long" indexed="true" stored="true" multiValued="false" />
<field name="id" type="string" indexed="true" stored="true" required="true" multiValued="false" />
<field name="ip" type="text_general" indexed="true" stored="true"/>
<field name="request_date" type="date" indexed="true" stored="true"/>
<field name="request" type="text_general" indexed="true" stored="true"/>
<field name="department" type="string" indexed="true" stored="true" multiValued="false"/> <field
name="category" type="string" indexed="true" stored="true" multiValued="false"/>
<field name="product" type="string" indexed="true" stored="true" multiValued="false"/>
<field name="action" type="string" indexed="true" stored="true" multiValued="false"/>
…
```

Create your Index cont.

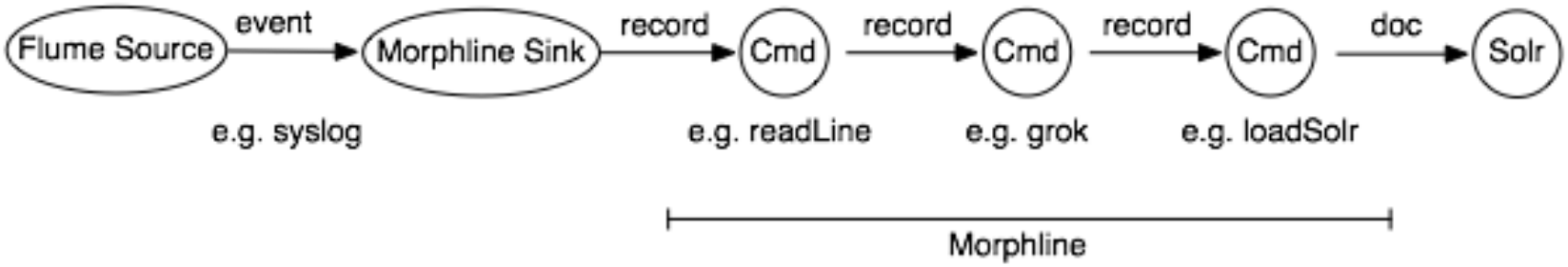
- Upload your configuration for a collection to ZooKeeper

```
$ solrctl --zk  <ALL YOUR ZK IPs>/solr instancedir --create live_logs
./live_logs_dir
```

- Tell Solr to start serving up a collection and start indexing data for it

```
$ solrctl --zk <ALL YOUR ZK IPs>/ solr collection --create live_logs -s 2
```

Flume and Morphline Pipeline



Flume with Morphlines Configured

- Easy to create custom Morphlines too…

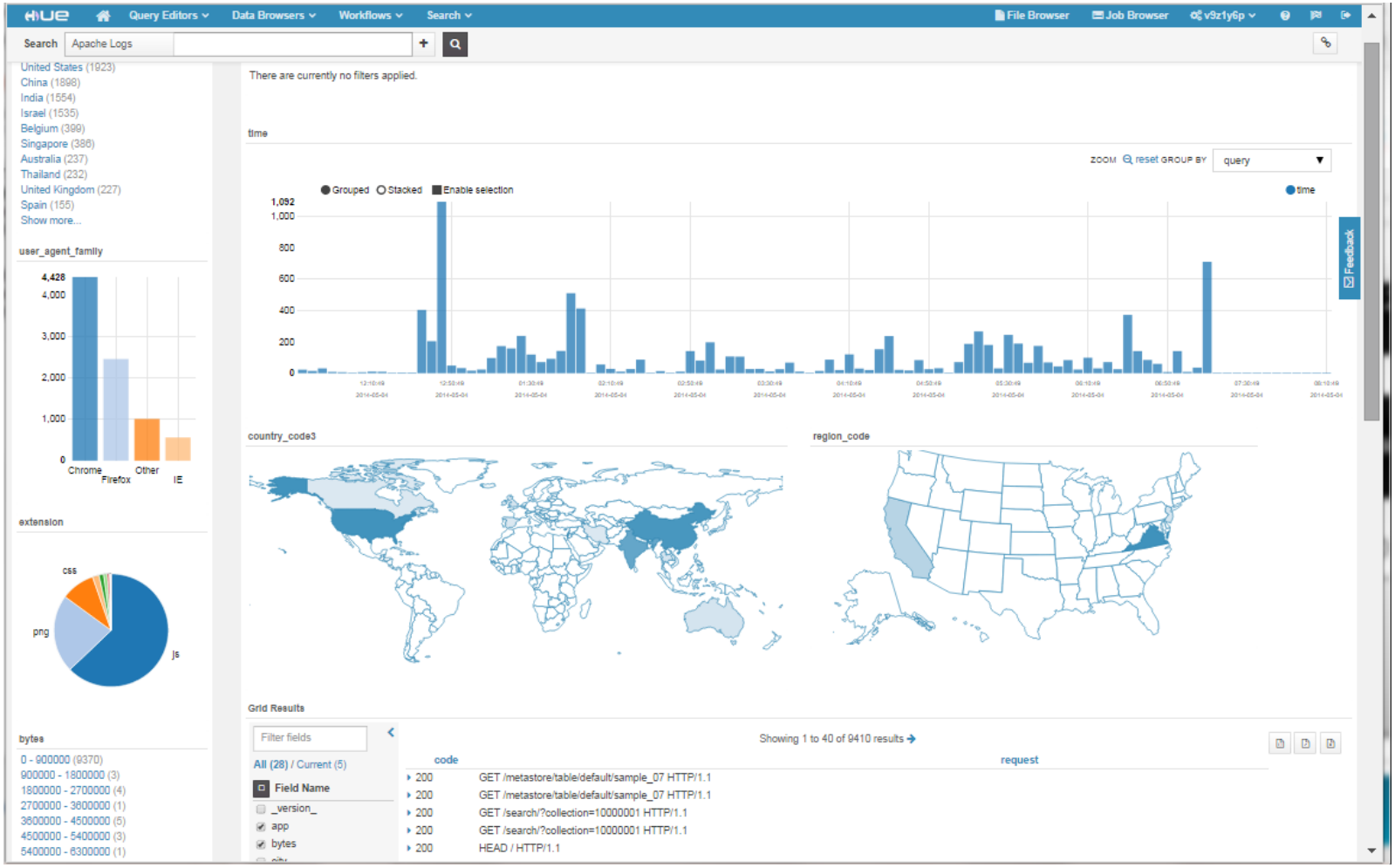
```
...
Pattern pCategory = Pattern.compile("/department/(.+?)/category/(.*)");
Matcher mCategory = pCategory.matcher(request_key);

while (mCategory.find())
{
    department = mCategory.group(1);
    category = mCategory.group(2);
    action = "view category products";
}
…
```

- Configure Flume to use your Morphlines and post parsed data to Solr

```
….
# Describe solrSink agent1.sinks.solrSink.type = org.apache.flume.sink.solr.morphline.MorphlineSolrSink
agent1.sinks.solrSink.channel = memoryChannel agent1.sinks.solrSink.batchSize = 1000
agent1.sinks.solrSink.batchDurationMillis = 1000 agent1.sinks.solrSink.morphlineFile =
/opt/examples/flume/conf/morphline.conf agent1.sinks.solrSink.morphlineId = morphline
agent1.sinks.solrSink.threadCount = 1
…..
```

Design your Search UI in Hue



Want to try Yourself?

- Try Cloudera Live (post 10/6)
  - Free mini-clusters to explore
  - Self-guided tutorials and code examples
  - Find more info (soon) at: **cloudera.com/live**
- For now
  - Play with read-only **demo.gethue.com**

Takeaways

- Information driven business is key forward
- Hadoop et al is a powerful technology ecosystem
  - Enables Enterprise Data Hub architecture
  - Addresses various big data challenges
- Use the right tool for the right workload
  - They are all conveniently available in the same platform
- Everybody can gain from Big Data principles!
  - Do the same workloads, but over larger data sets
  - Gain more insight by using multiple data sets to serve business questions
  - Cost-efficiently serve multiple use cases over same data via an EDH architecture
  - Much easier to change your mind…

Did you learn something?

Don’t forget to VOTE!!!

Q&A

- Learn more
  - Cloudera University
    - training, certification, free on-line classes
  - Join the Community
    - dev2dev forums, community email lists, HUGs, …
- Reach me
  - @EvaAndreasson
- After the break
  - Part 3 with Dean Wampler – woot!!


cloudera®
Ask Bigger Questions

Common Use Cases

- Threat detection
- Active archive / accessible global knowledge base
- Data accuracy
- Streamlined cross-data type aggregation
- Richer customer profiling / ecommerce experience
- Interactive market segmenting / customer identification
- Expedited data modeling
- ….