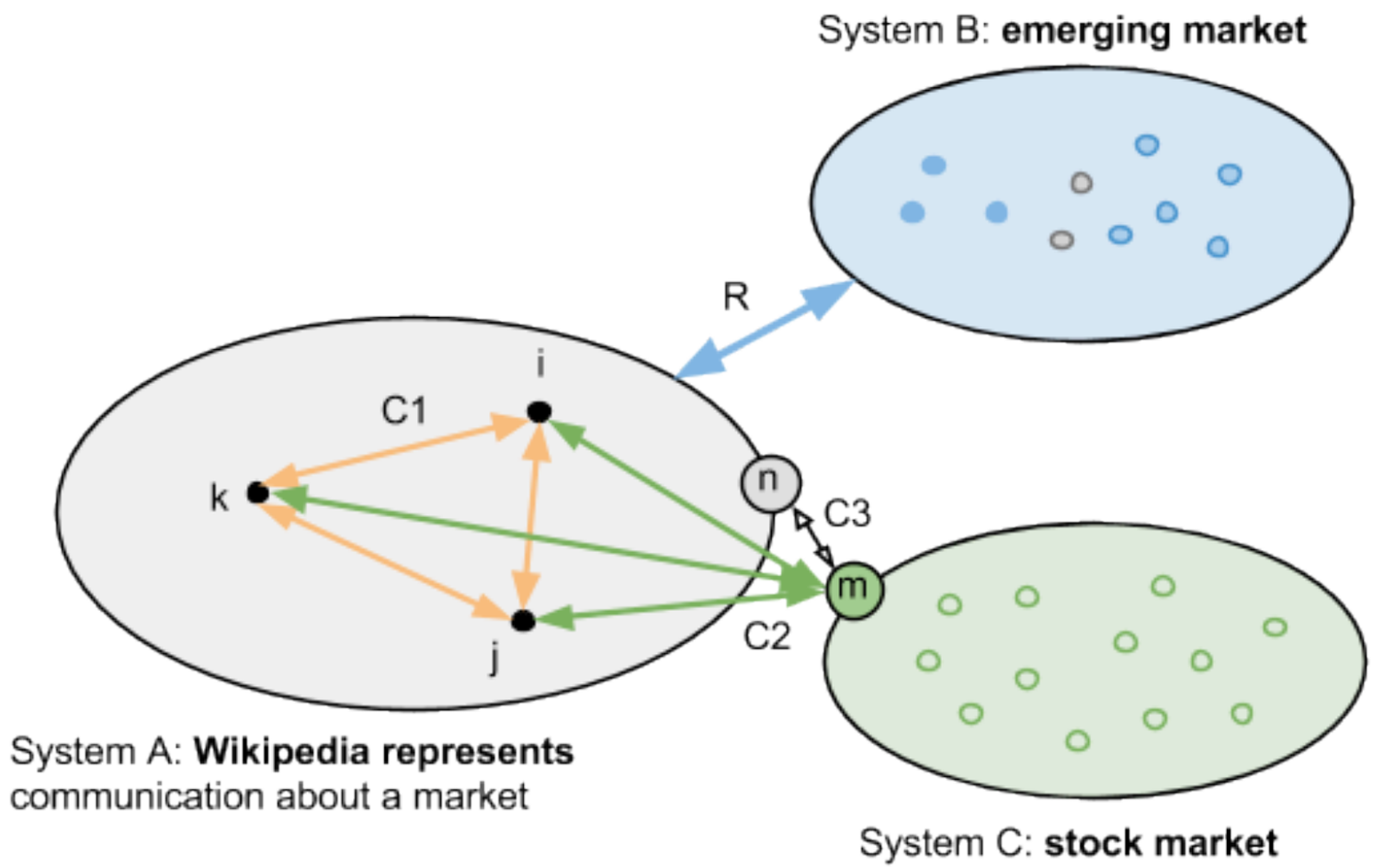
Managing Complexity

About using many datasets in multiple Hadoop clusters

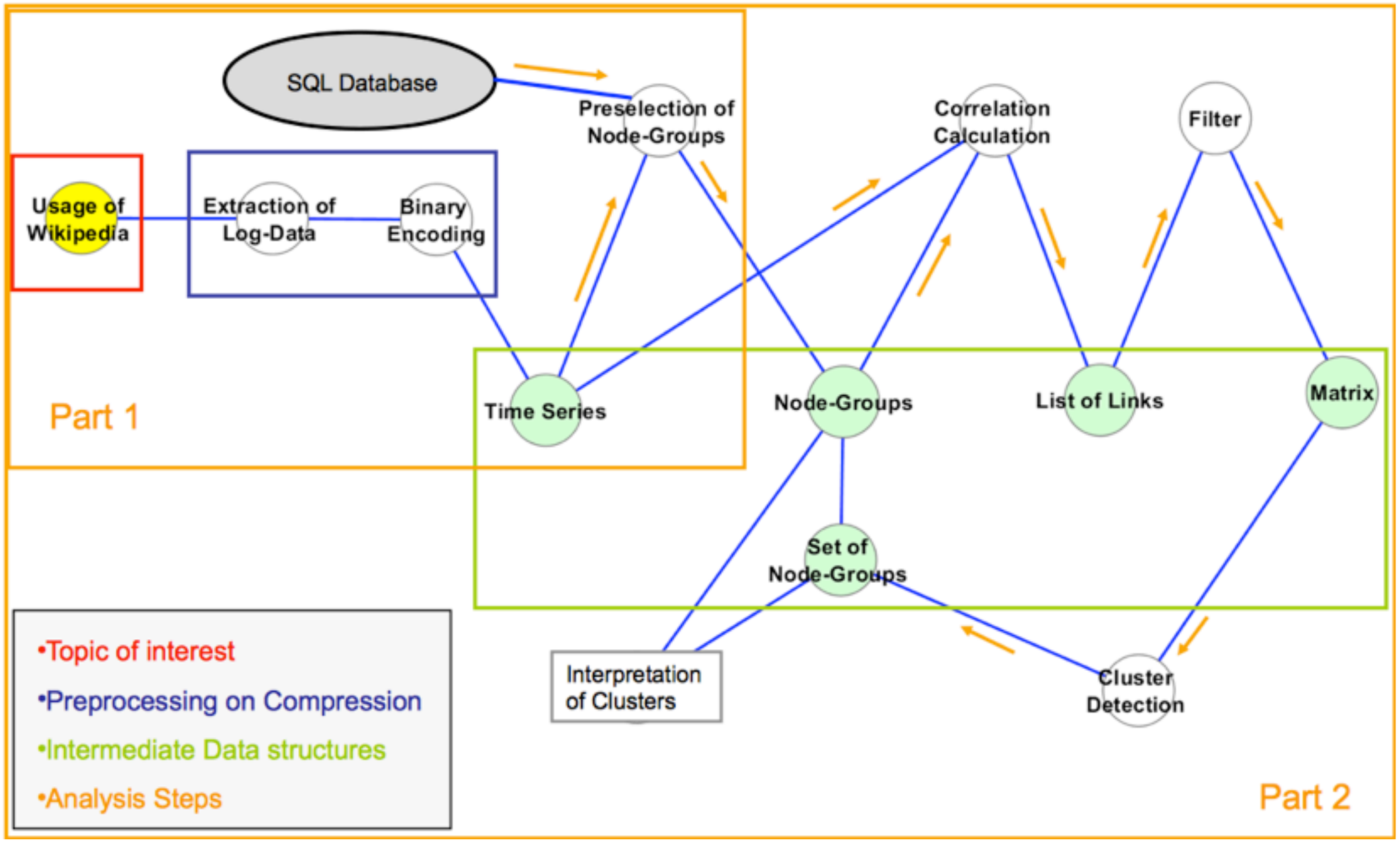
- Use Case Overview
- Lessons Learned
- Project Proposal: ETOSHA

Mirko Kämpf, Q12014, Cloudera Inc.

Data Driven Market Studies

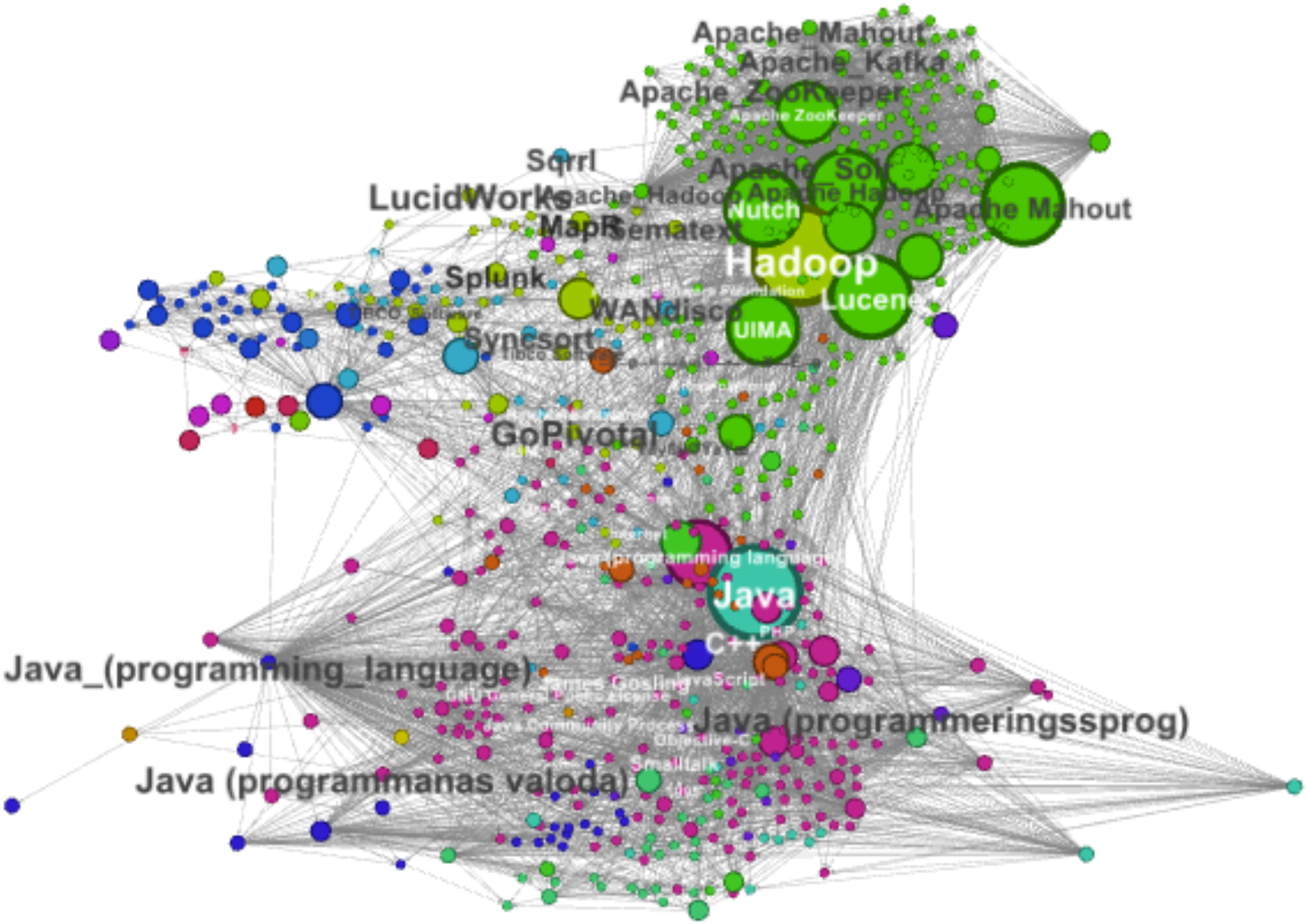


Analysis Scenario ...

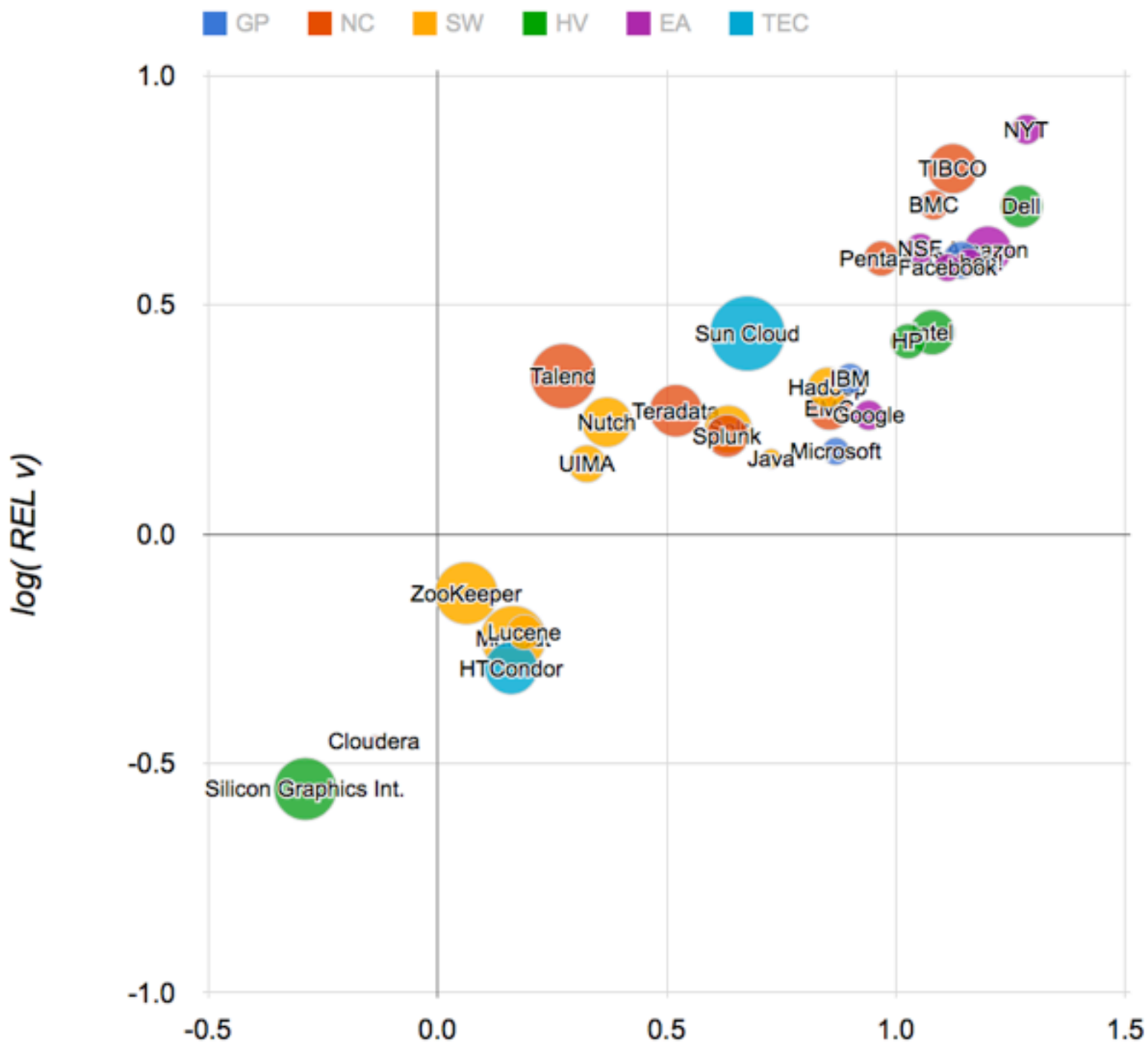


Social Media Analysis - more than a buzzword! - Based on daily access rates to Wikipedia pages (or even groups in a given semantic context) one can study complex systems like financial markets or technology evolution and emerging markets, like the market around the Hadoop Ecosystsem.

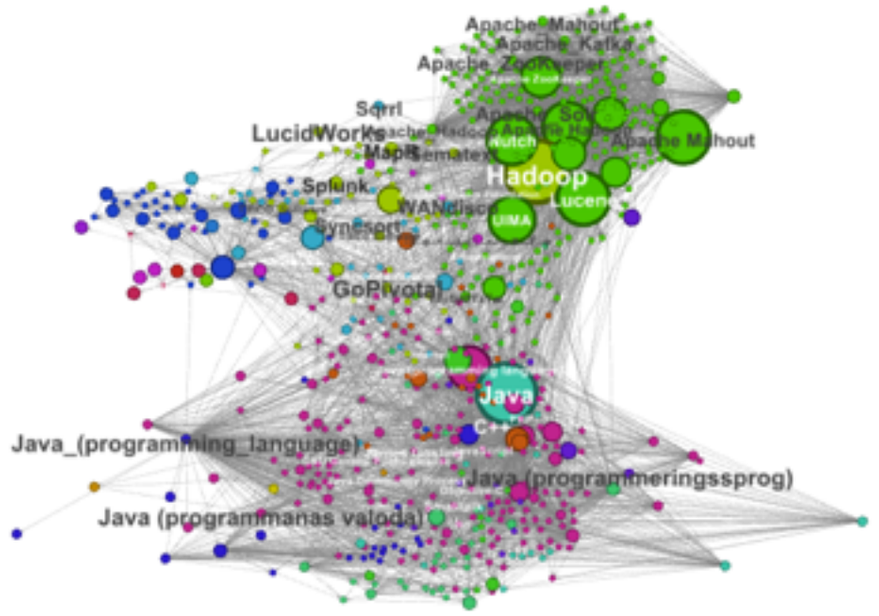
Hadoop: An emerging market?



First results of a case study:

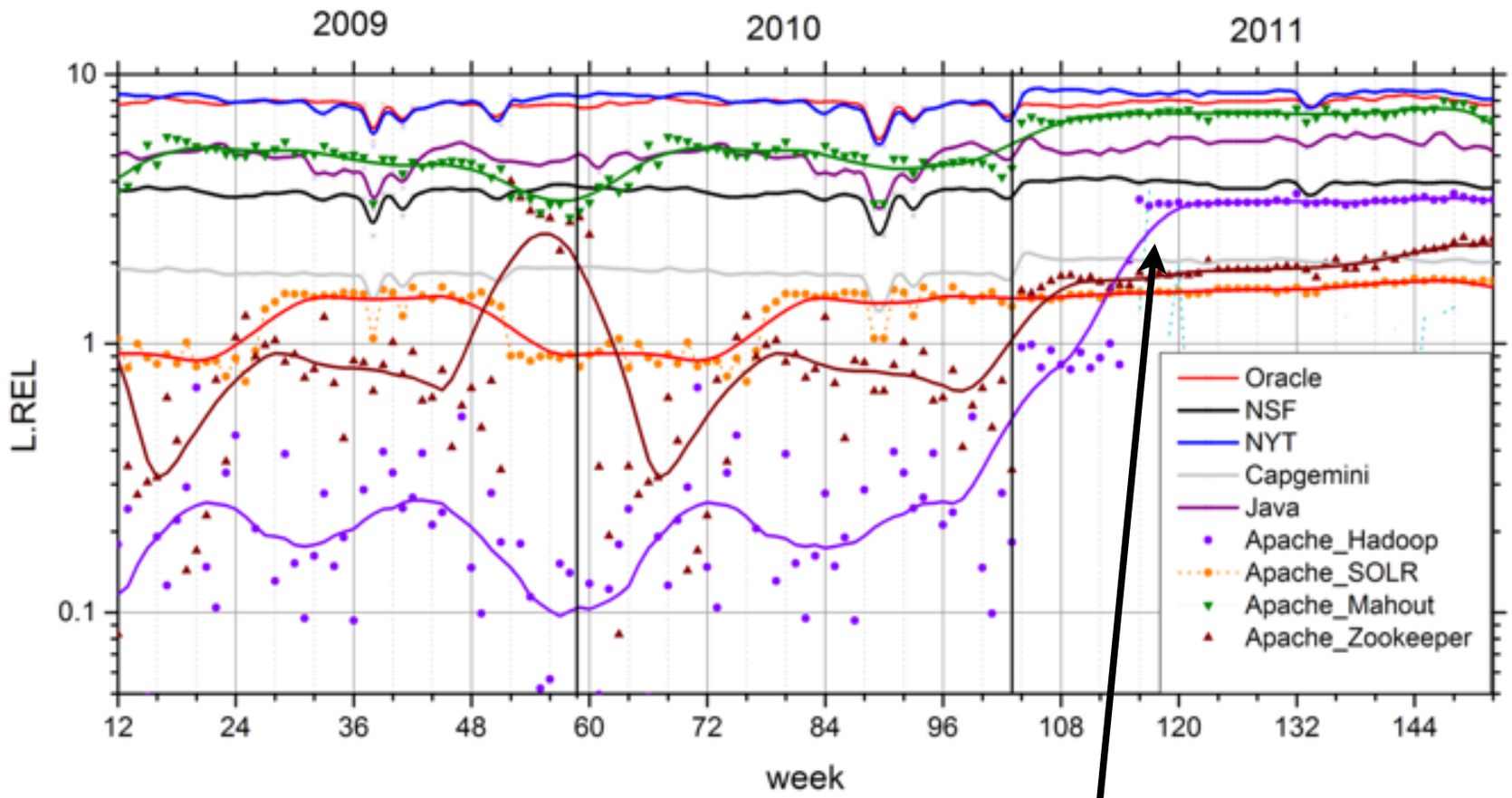


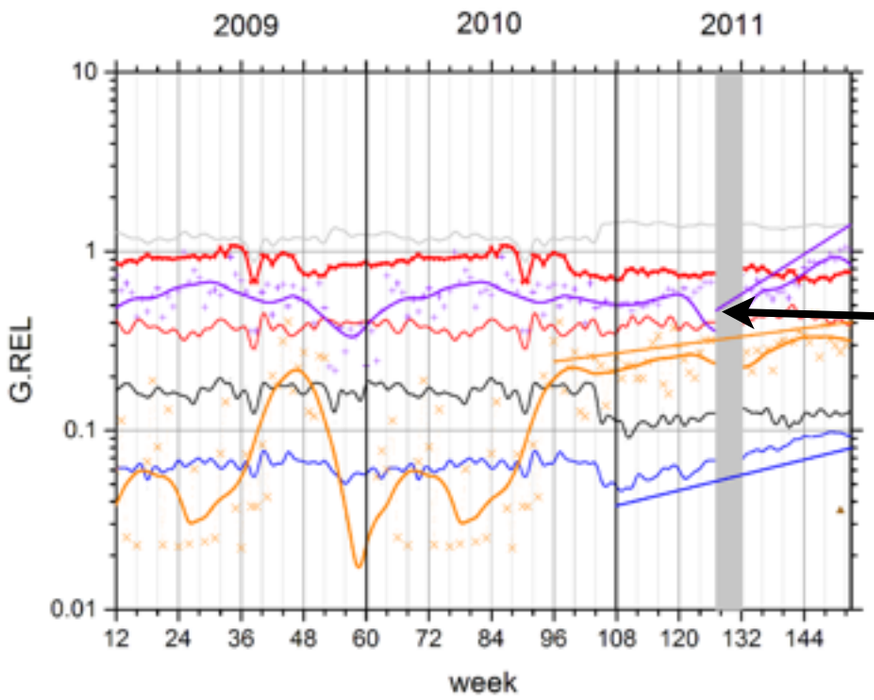
relevance index



network of pages

Public recognition ... local vs. global

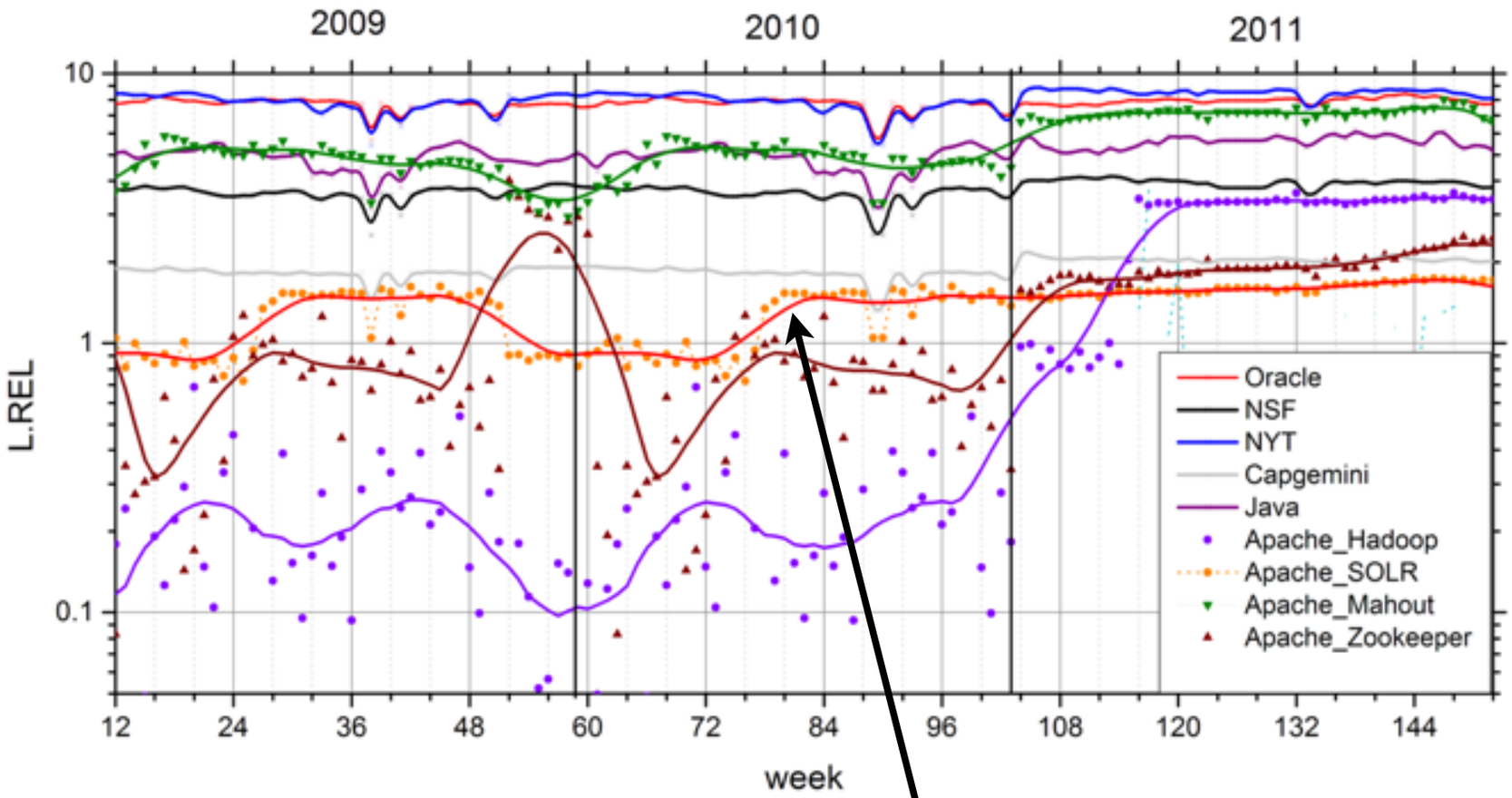


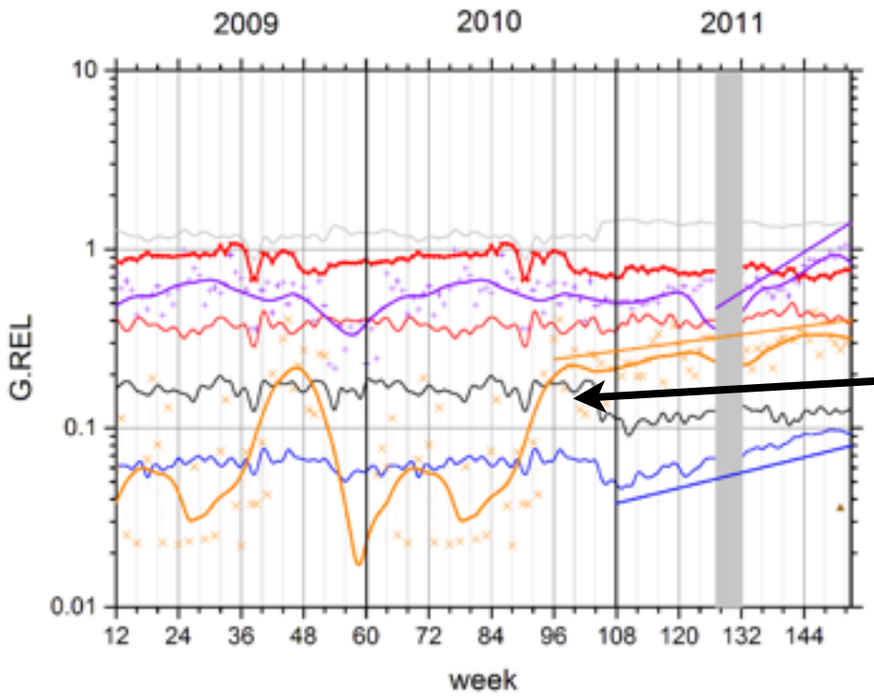


Q1: 2011 hadoop reaches a stable relevance level in **english** language

Q2: 2011 hadoop's relevance starts growing in all other languages with a delay of about 6 months.

Public recognition ... local vs. global





Q2: 2010 SOLR reaches a stable, and still slowly growing relevance level above 1 in **english** language.

Q3: 2010 SOLR's relevance remains stable slightly higher level and is continuosly growing in all other languages with a delay of about 4 months.

First Step: (Re)Thinking the Process ...

We need stable, repeatable, and traceable processes with high quality process documentation.

My code is my documentation, might work well in simple MapReduce but how do I track all my Hive queries, which have been executed with **flexible parameters**?

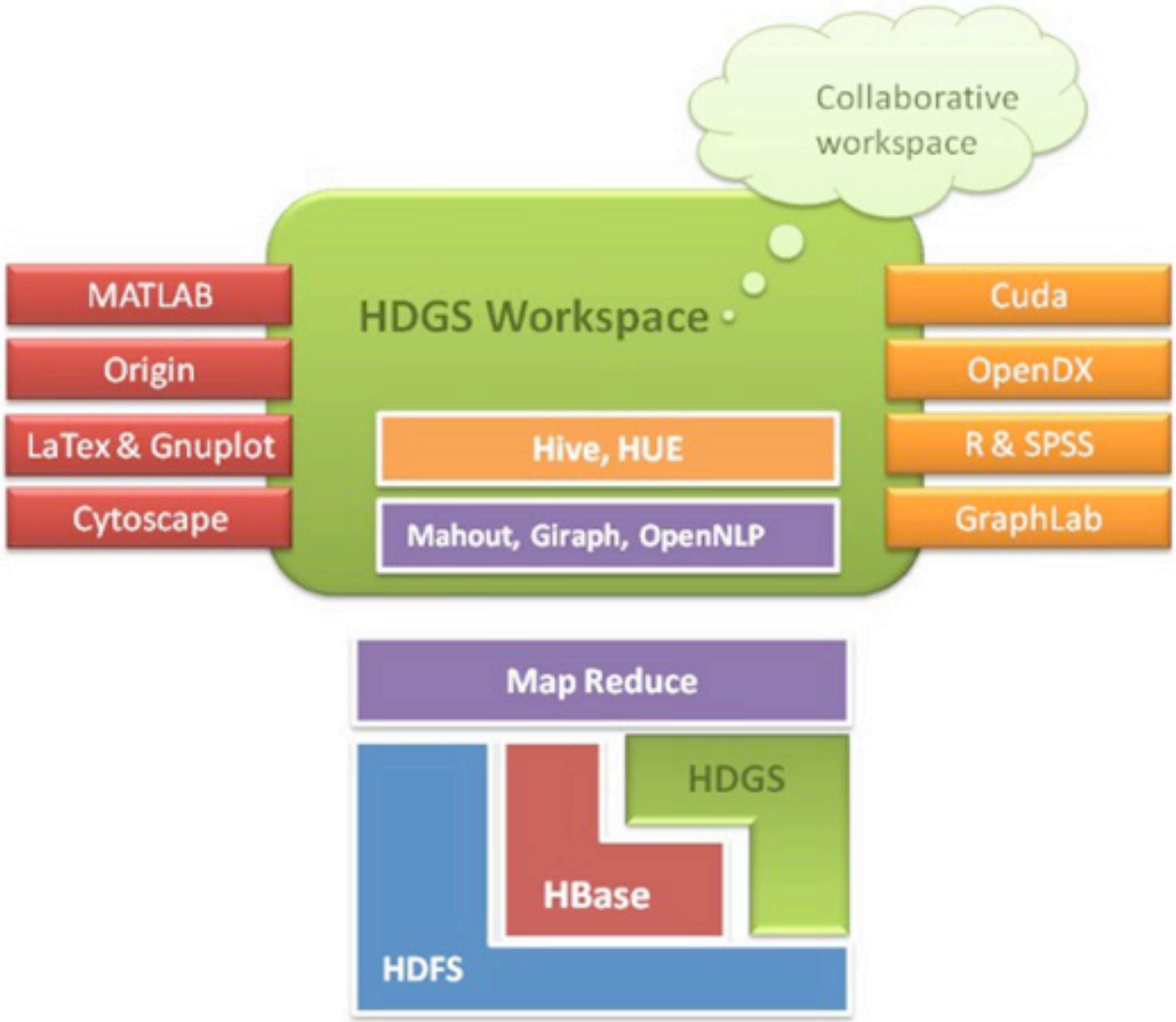
Second Step: Care about Metadata ...

We need stable, repeatable, and traceable processes
with high quality process documentation.

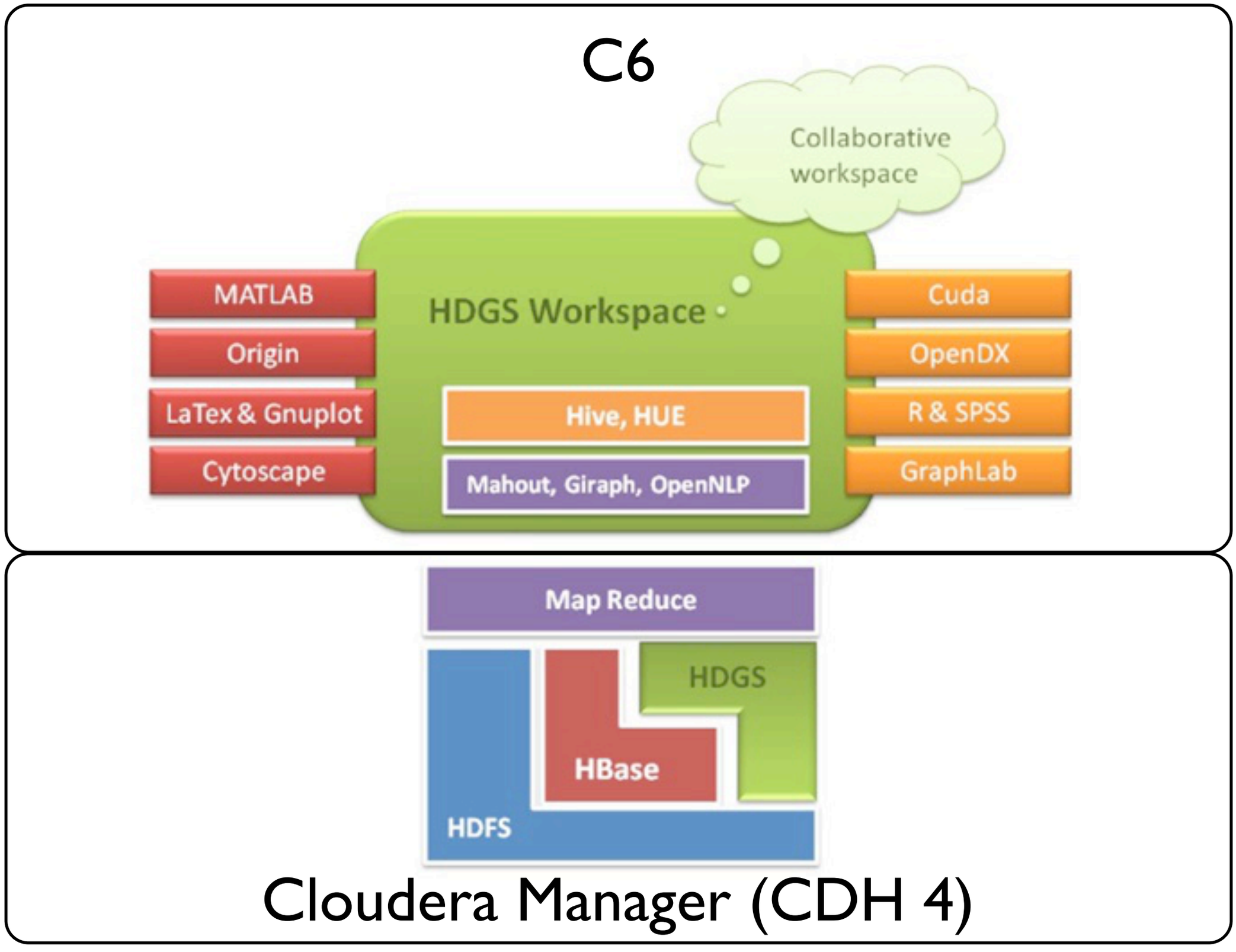
My code is my **documentation,** might work well in simple MapReduce
but how do I track all my Hive queries, which have been executed
with **flexible parameters**?

What data is where and how was it (pre)processed?
How about data quality, **is it worth** to run a job on it?

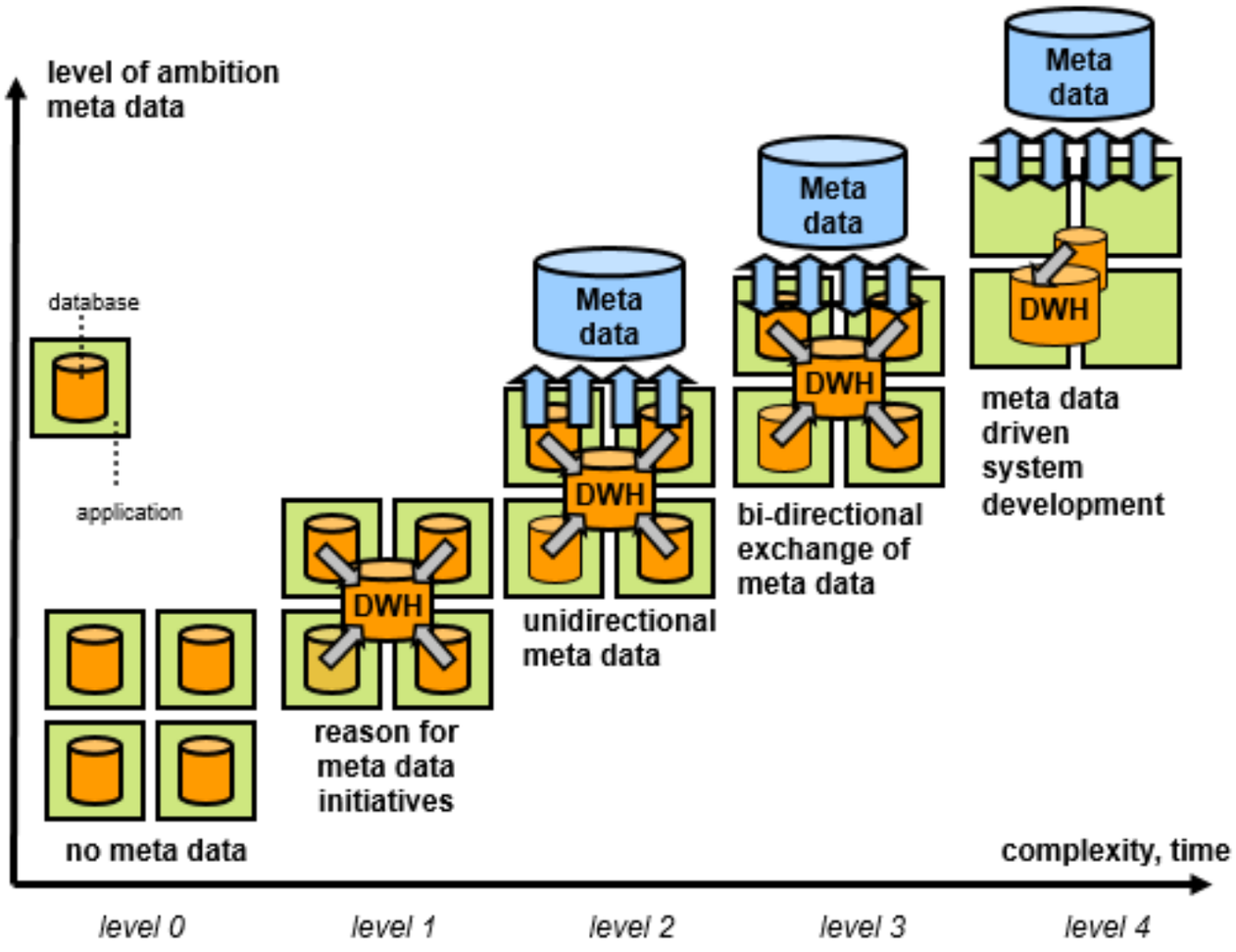
I have a dream ...



... it **requires:** distributed metadata



It's all about processes & metadata?



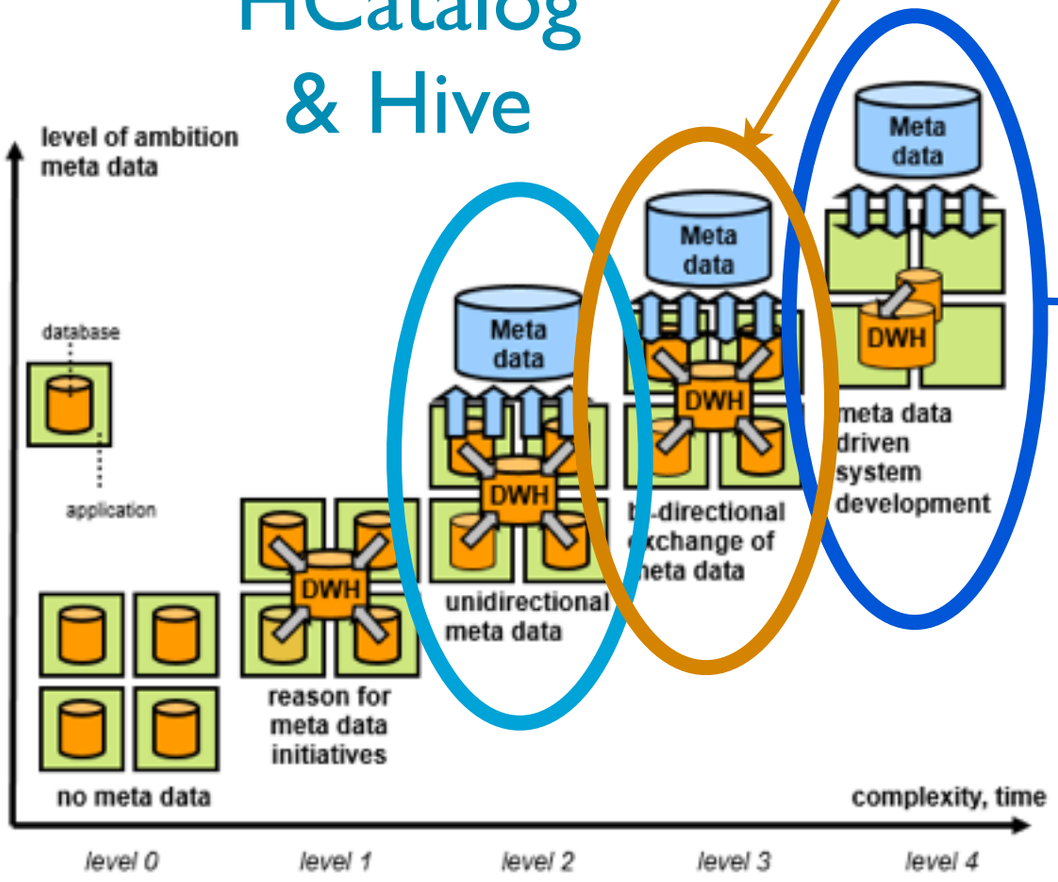
What data is where and how was it preprocessed?
How about data quality, is it worth to run a job on it?

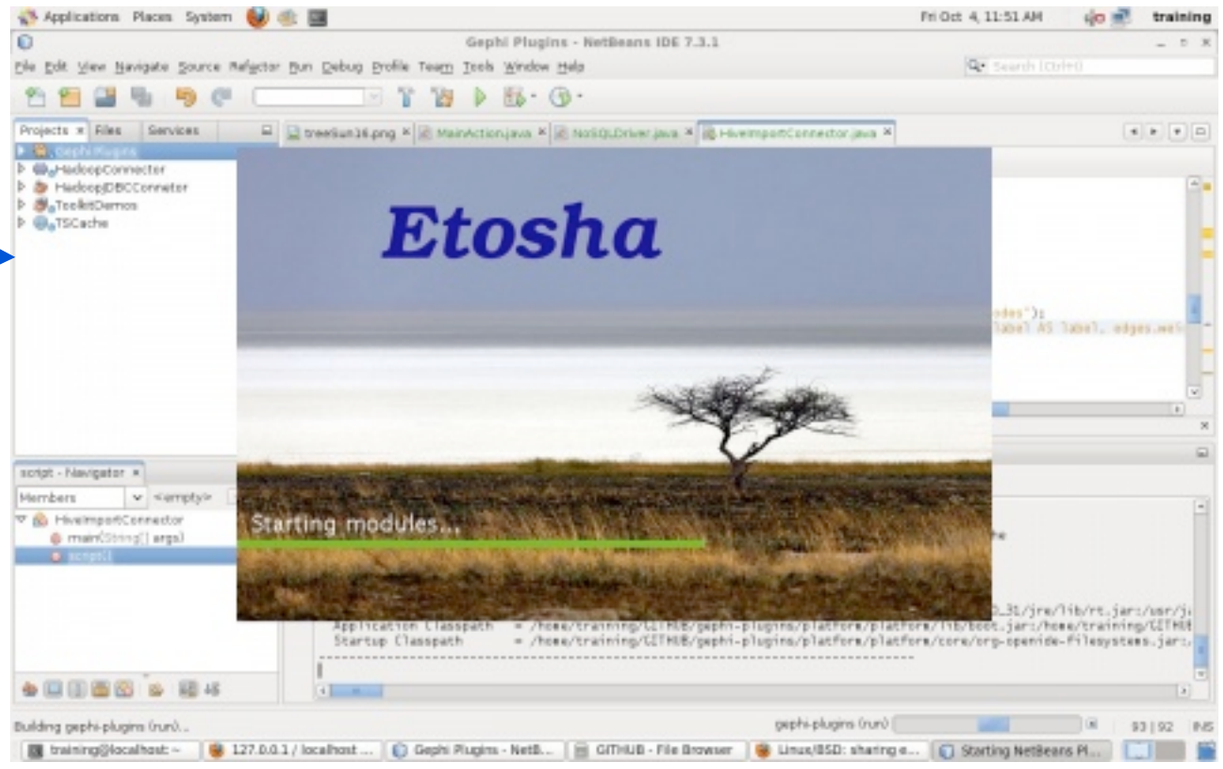
future:

ETOSHA

now:

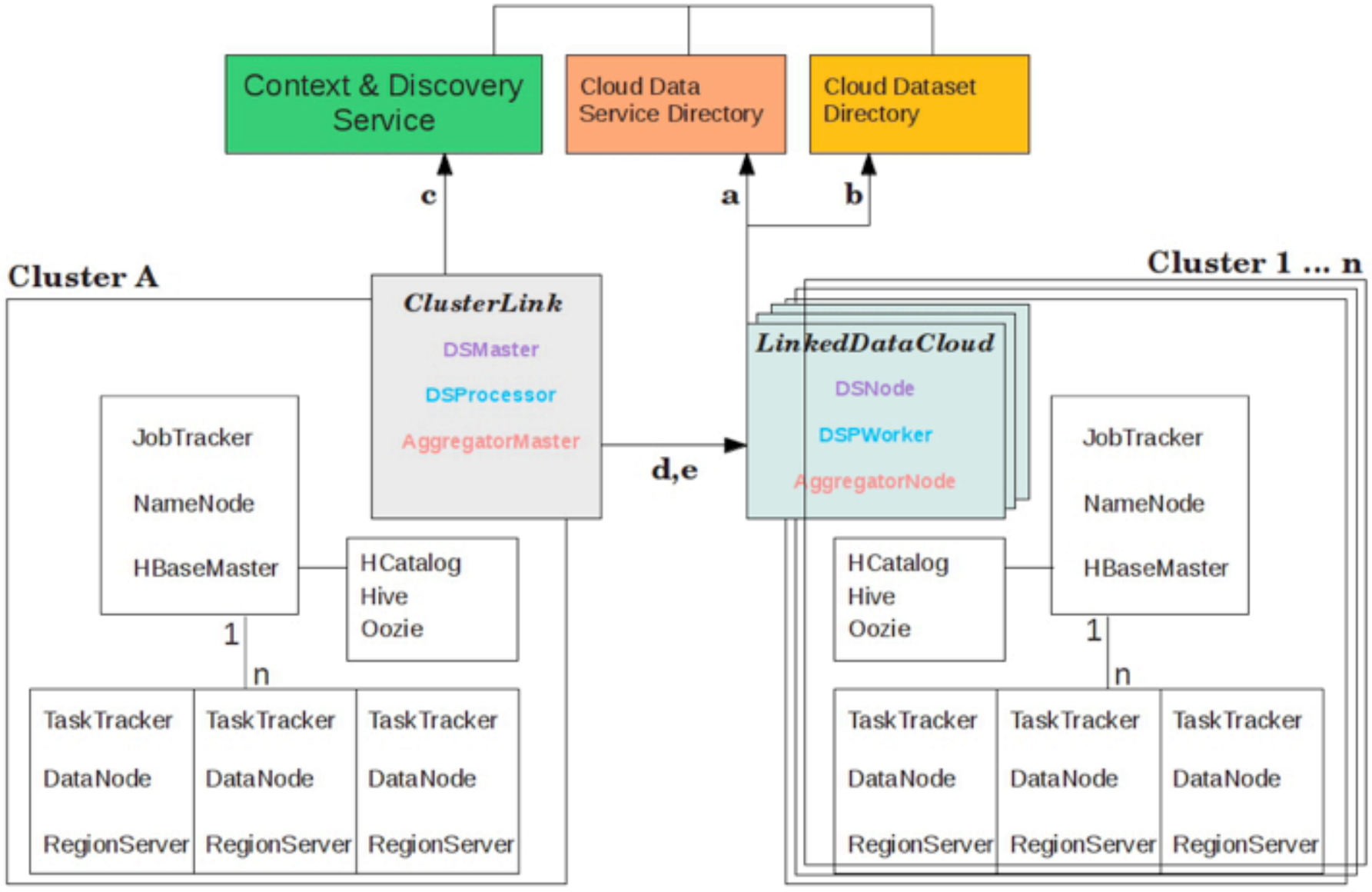
HCatalog & Hive





Architecture: **the bird's view**

Colored boxes represent required services and
can be implemented by arbitrary tools and applications.
They just have to implement the dataset integration framework.



Linked Data Cloud Masters - an data set integration layer (DSI-layer) on top of multiple Hadoop clusters works based on web services and semantic web technology the concept of ***data locality*** *will be achieved* ***across multiple clusters****. Cost tracking and optimization services are supportive for several business models following the* ***Data As a Service*** *paradigm.*

The path ...

MapReduce &
"inCluster **queries**"

MapReduce API

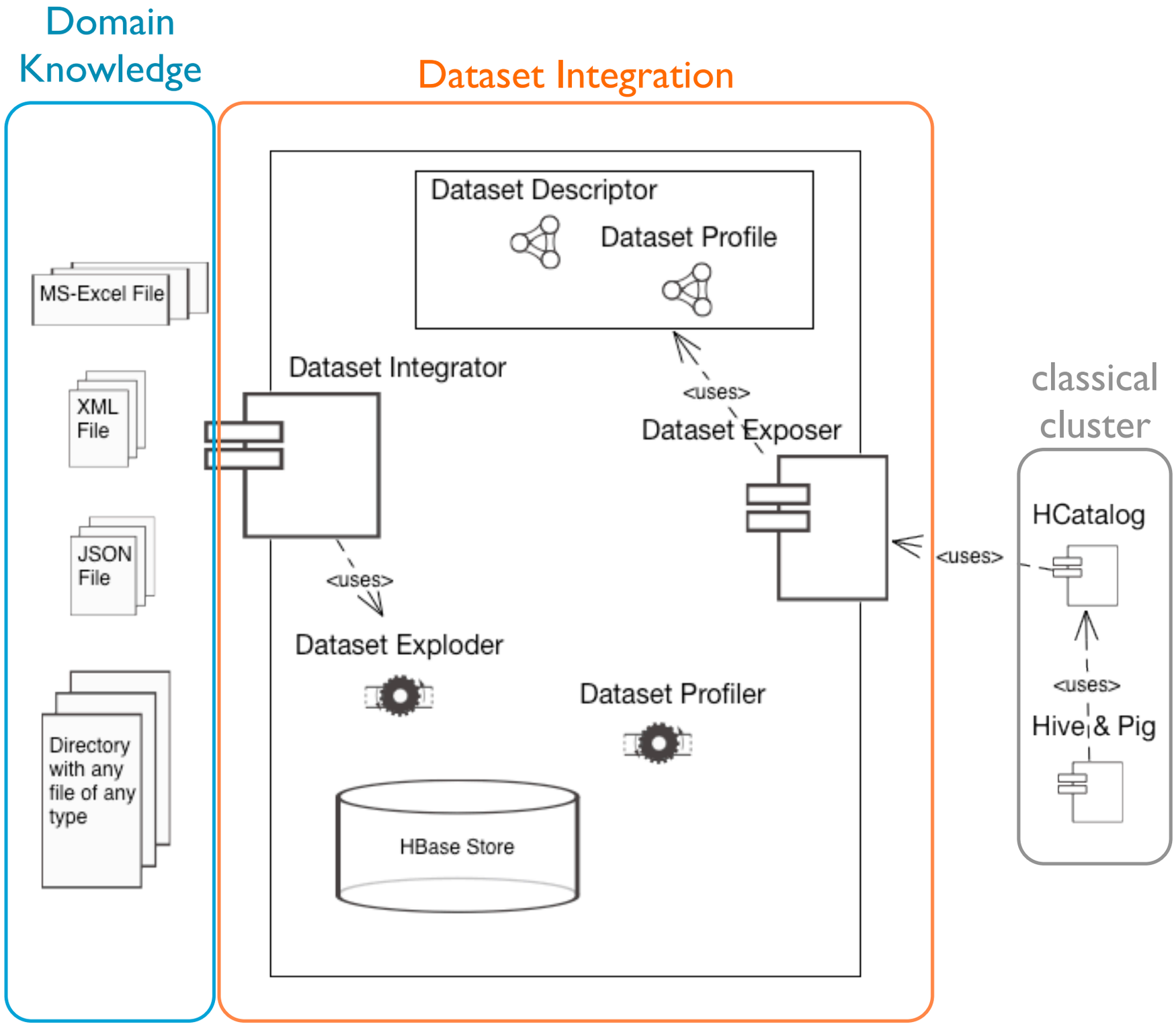
"inCluster **Workflows**"

Oozie, Sqoop, Flume

Cluster Spanning
Data Driven Business **&**
Research**, using linked datasets**

ETOSHA

Cluster Spanning Dataset Management



Two steps towards a prototpye ...

first results

- Job and Dataset Context
- Dataflow & -link Context

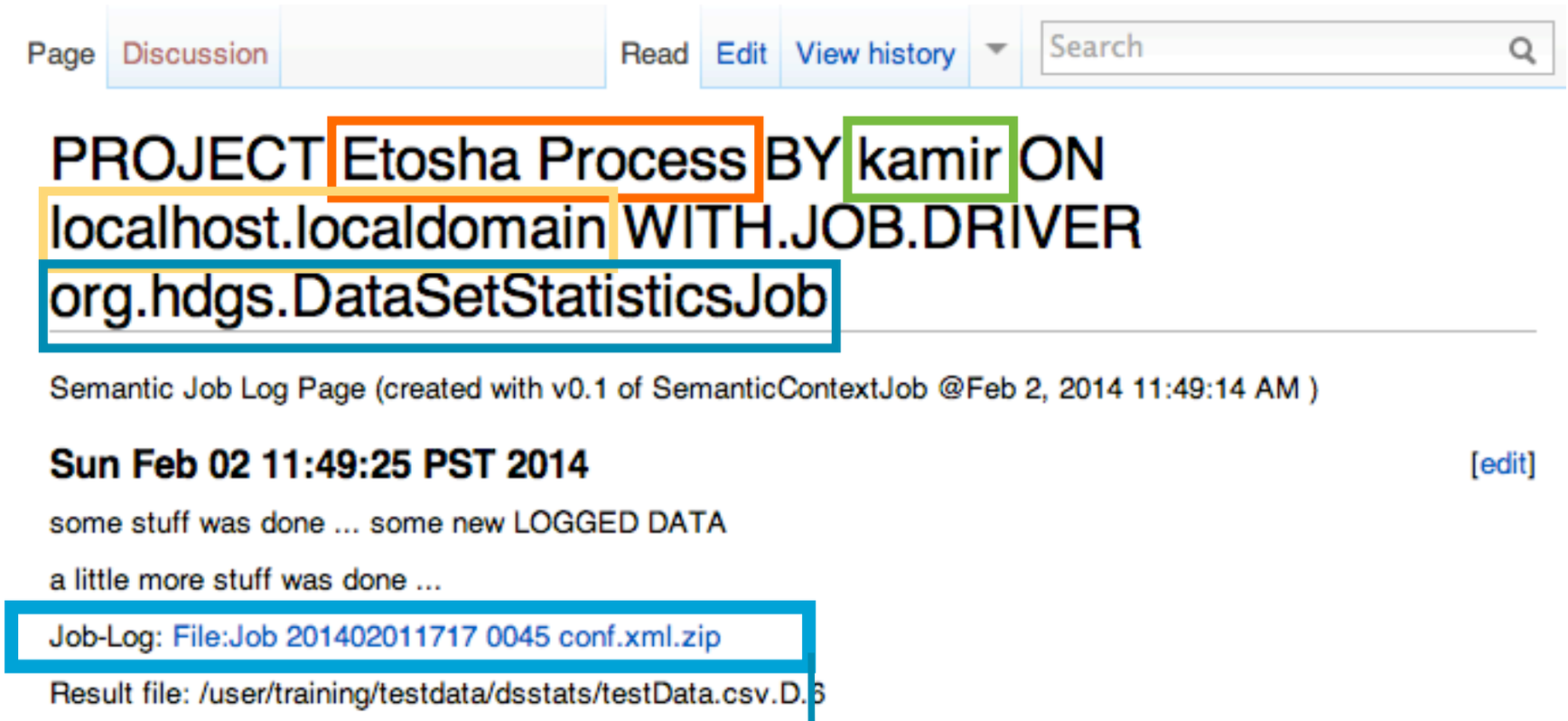
Step 1: Job and Dataset Context

- **Q:** What job did produce a dataset by using what algorithm specific parameters and at what cost?

  A: *Build in semantic logging collects metadata in a shareable knowledge base, which is searchable and has an API for tool integration.*

Transparency is achieved by traceability from final results **(charts, tables), down to the code including** runtime parameters **and** process logs.

Key concept: Wiki based knowledge graph with built in semantic annotations



Transparency is achieved by traceability from final results **(charts, tables), down to the code including** runtime parameters **and** process logs.

Sun Feb 02 12:05:36 PST 2014 [edit]

some stuff was done ... some new LOGGED DATA

a little more stuff was done ...

Job-Log: File:Job 201402011717 0047 conf.xml.zip

Result file: /user/training/testdata/dsstats/testData.csv.D.8

Dataset statistics: File:Job 201402011717 0047 conf.xml simple DS statistics.dat.zip

/user/training/testdata/dsstats/testData.csv.D.8

Result table:

0	1.0	25.0	25	325.0	13.0
1	2.0	50.0	25	650.0	26.0
2	4.0	100.0	25	1300.0	52.0
3	8.0	200.0	25	2600.0	104.0
4	16.0	400.0	25	5200.0	208.0

Categories: Etosha Process | Kamir | Localhost.localdomain | Org.hdgs.DataSetStatisticsJob | CDH4.2.VM.adht | Cloudera-Training-VM-4.2.1.adht@MacBookPRO

for each column of the data set we have a row of metadata with a well defined meaning, encoded in an ontology

Metadata processors generate searchable / queryable results

Transparency is achieved by traceability from final results **(charts, tables), down to the code including** runtime parameters **and** process logs.

Step 2: **Dataflow & -link Context**

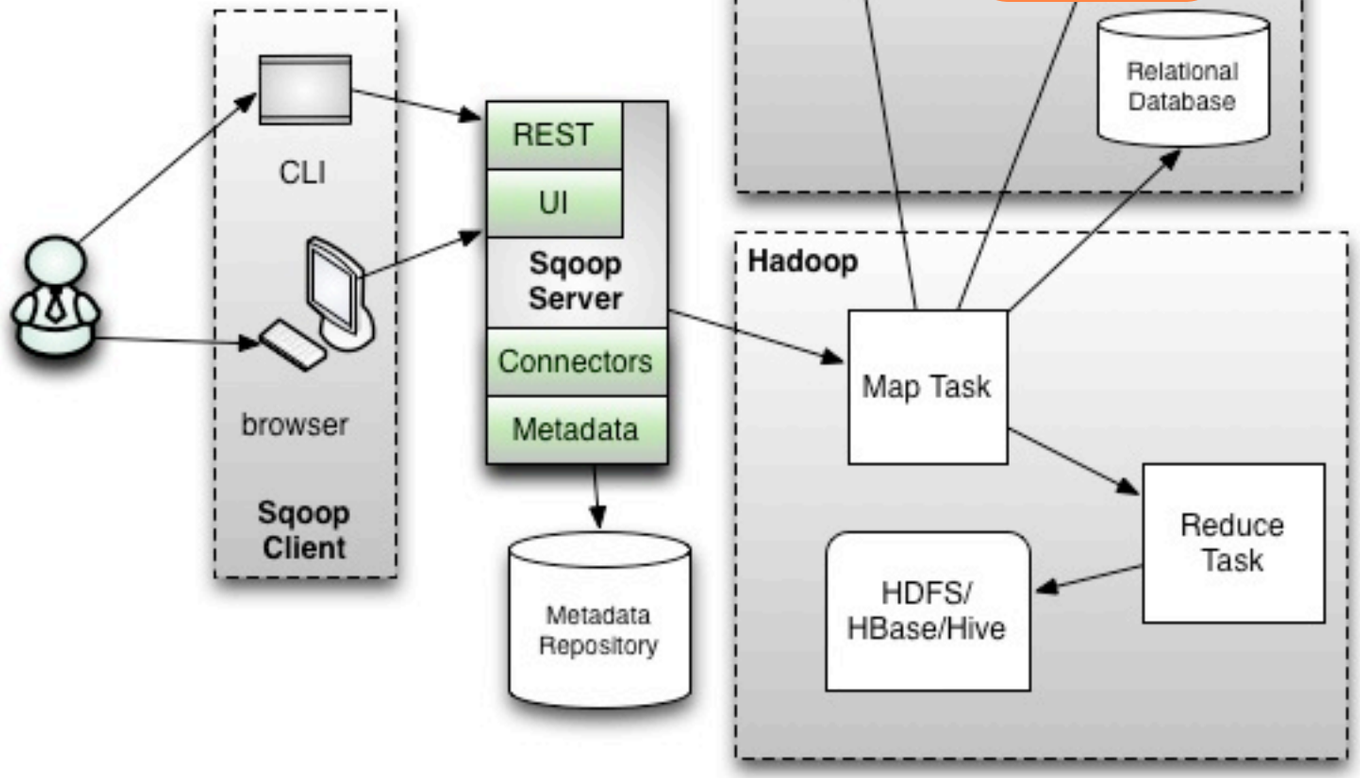
- **Q: How can I join data from multiple Office documents without the manual export nightmare?**
  A: *Dataset integrators map content from std. Office-file types to tables in Hadoop, and allow fast access via Impala, or batch jobs via Hive and MapReduce.*

Domain focused flexibility **is achieved by** semantic mapping definitions**, which allow any kind of data extraction and migration, based on established technologies like Flume, Sqoop, JMS, and XML.**

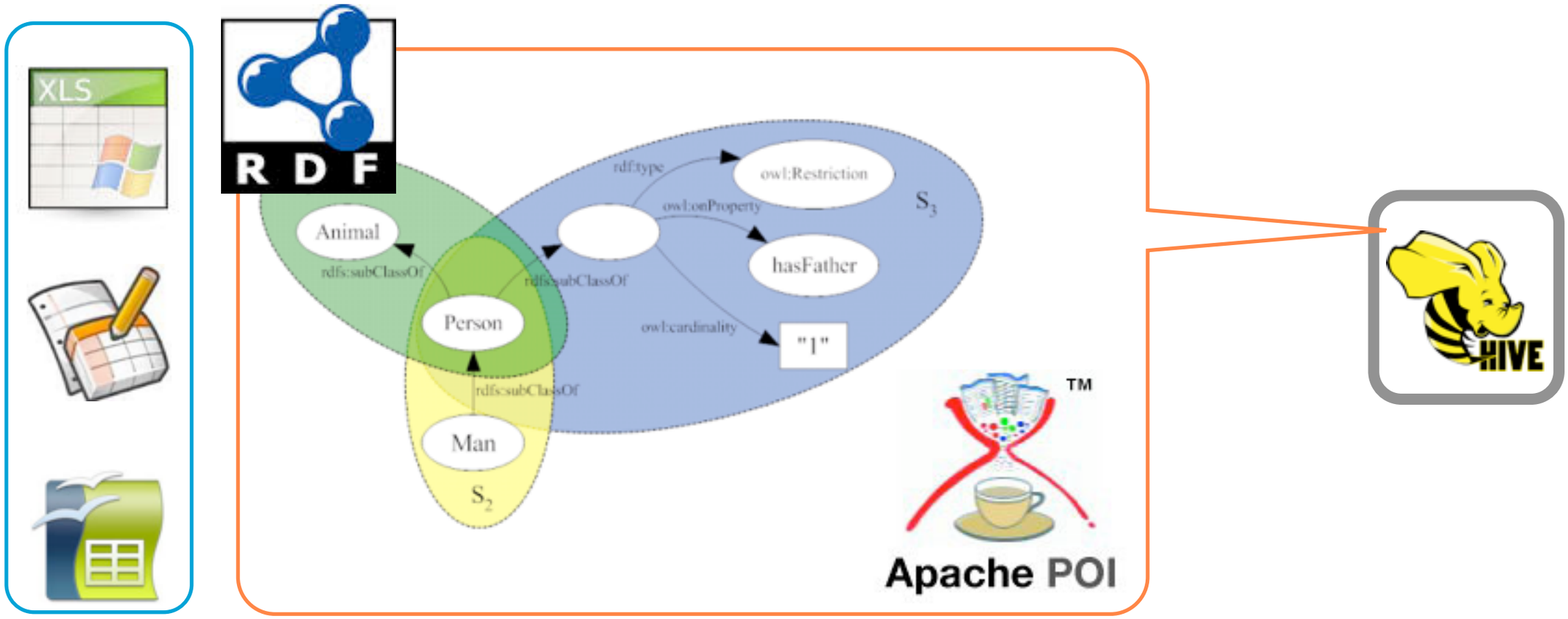
Step 2: Dataflow & link Context

Domain Specific Documents

key concept: semantic mapping



Step 2: Dataflow & link Context



Domain focused flexibility is achieved by **semantic mapping definitions**, which allow any kind of data extraction and migration, based on established technologies like Flume, Sqoop, JMS, and XML.

*Can also be done via Oracle SOA Suite 11g

Next Steps:

- **Release 0.1** April 2013)
  - **finish the generic context-log framework** (reference and sample apps)
  - **finish the doc-mapper cartridge framework**
  - integrate the whole app in HUE
- **Release 0.2** October 2013)
  - **finish the Dataset-Exploder and Dataset-Profiler**
  - integrate doc-mapper cartridges in to Sqoop
- **Release 0.3** $t_{release}$ = **f**(resources, demand)
  - finish the cluster spanning dataset integration layer which uses a shared semantic knowledge graph

any feedback welcome: mirko@cloudera.com