APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Big Telco, Bigger DW Demands: Moving Towards SQL-on-Hadoop

Presented For The Apache Foundation By
LINUX FOUNDATION

Keuntae Park

SK telecom

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

- IT Manager of SK Telecom, South Korea's largest wireless communications provider

- Work on commercial products (~'12)
  - T-FS: Distributed File System
  - Windows compatible layer on TimOS
  - T-MR: on-demand MapReduce service like E-MR
- Open source activity ('13~)
  - Committer of Apache Tajo project

TAJO

Overview

- Background
  - Telco requirements
- Before Tajo
  - Commercial product
  - Open source (Hadoop) outsourcing
- After Tajo
  - Issues & solutions
  - Performance
- win-win between community and company
- Future Works

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Telco data characteristics

- Huge amount of data
  - 40 TB/day (compressed)
  - 15 PB (estimated, end of 2014)
- Report & OLAP ad-hoc query
  - Filtering
  - Summary
  - BI tools

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

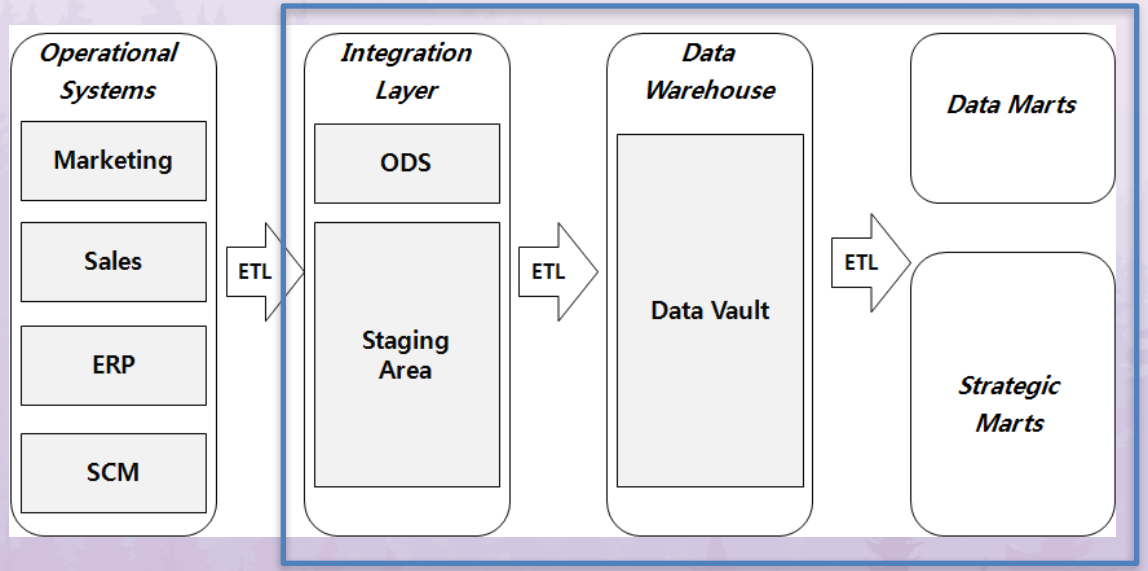
Presented For The Apache Foundation By
LINUX FOUNDATION

Requirements - different size, different speed

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

	Filtering & aggregation	Summary	Data re-construction	BI report	Ad-hoc Query
Target	accumulated for 5 minutes	daily sum of filtered data	entire summary data	mart data	summary data
Frequency	every 5 minutes	daily or monthly	non-regularly (rare)	ah-hoc	ah-hoc
Amount of data	terabytes	hundreds of terabytes	petabytes	tens of gigabytes	tens of terabytes
Response time	within a minute	within a hour	no strict deadline	within two seconds	within a hour

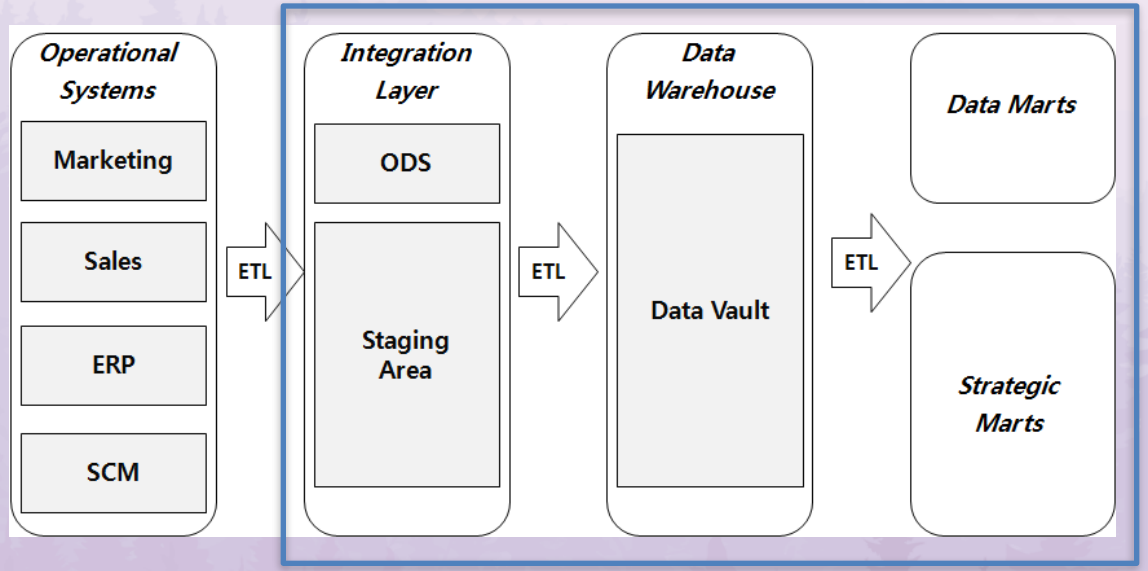
Previous approach - DBMS



based on MPP DBMS

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Previous approach - DBMS



based on MPP DBMS

Too Expensive

Not Scalable

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Previous approach - DBMS

Too Expensive

Not Scalable

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

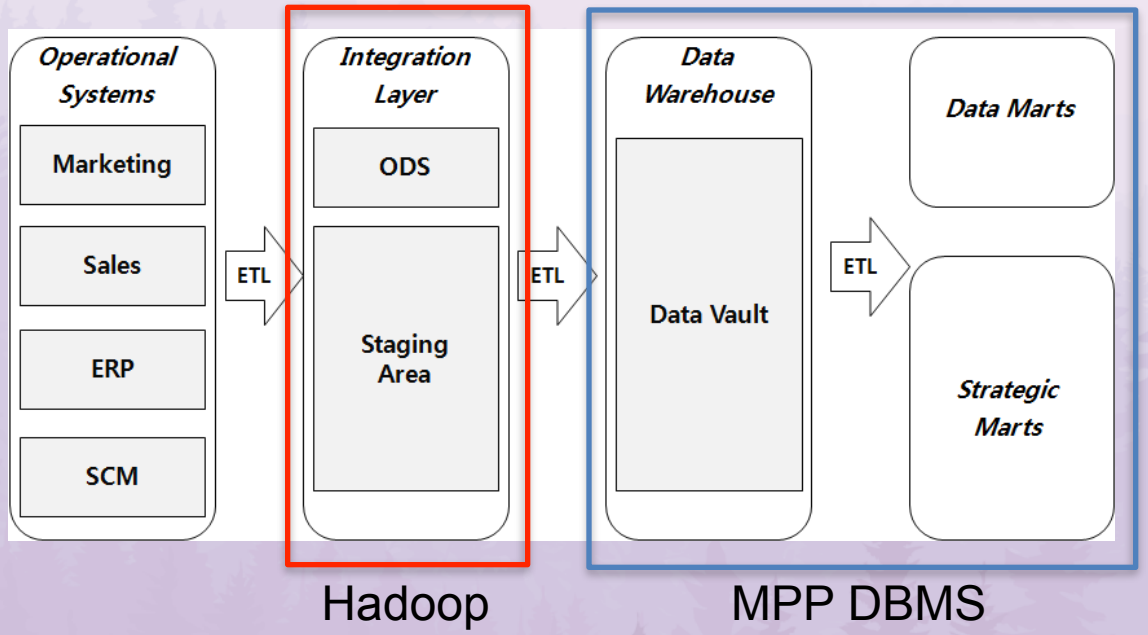
Previous approach - DBMS

sive

ble

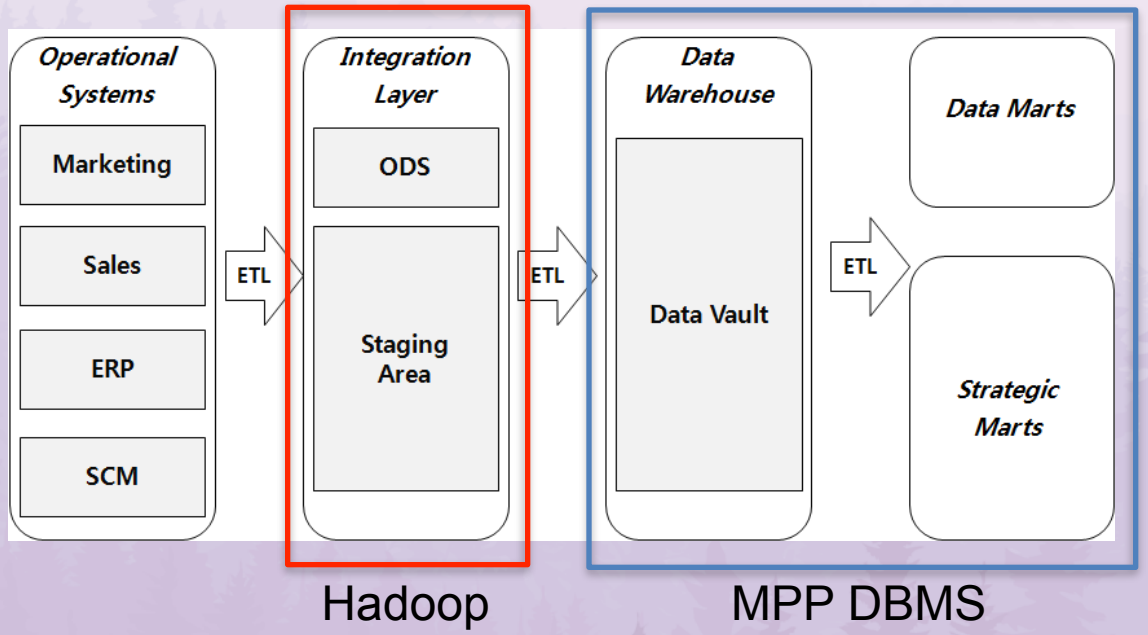
based on MPP DBMS

Previous approach - Hadoop(MapReduce, Hive) + DBMS



APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Previous approach - Hadoop(MapReduce, Hive) + DBMS



Working (but…)

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Still has Problems

- Hadoop outsourcing
  - quality of outcome is not good (actually bad)
  - communication overhead
  - hard to reflect requirements on open source
- Data Warehouse and Mart becomes bigger

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Solution - Tajo!!

- It can replace both DBMS and Hadoop
  - High throughput for batch processing
  - Low latency for ad-hoc queries
  - ANSI SQL compatible
- Can do by myself
  - very open community
    - easily make issues about what I really need
  - fast growing
    - issues solved very fast

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

About Tajo

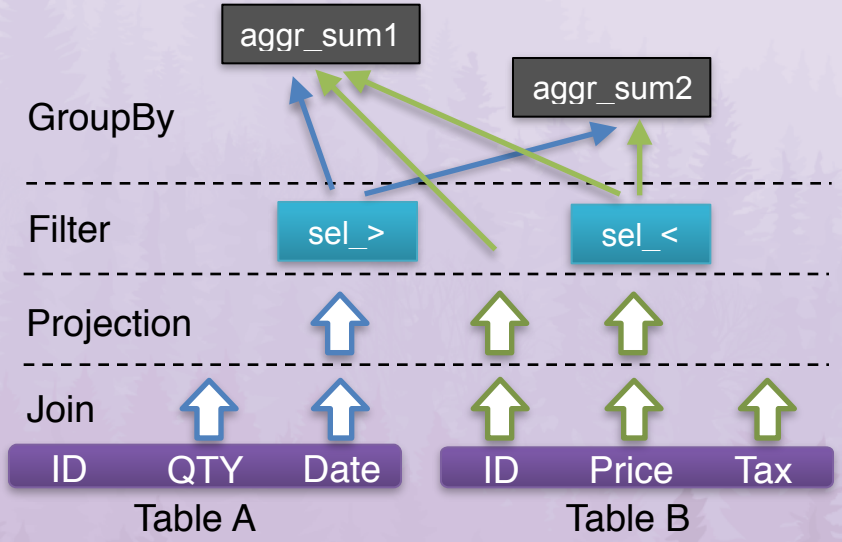
- Tajo (since 2010)
  - Big Data Warehouse System on Hadoop
  - Apache top-level project (entered the ASF in March 2013)
- Features
  - SQL standard compliance
  - Fully distributed SQL query processing
  - HDFS as a primary storage
  - Relational model (will be extended to nested model in the future)
  - ETL as well as low-latency relational query processing (100 ms ~)
- News
  - 0.2-incubating: released November 2013
  - graduation to top-level: April 2014

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo logical optimizer

- Cost-based join ordering
- Projection/Filter push down & Duplicated expression removal



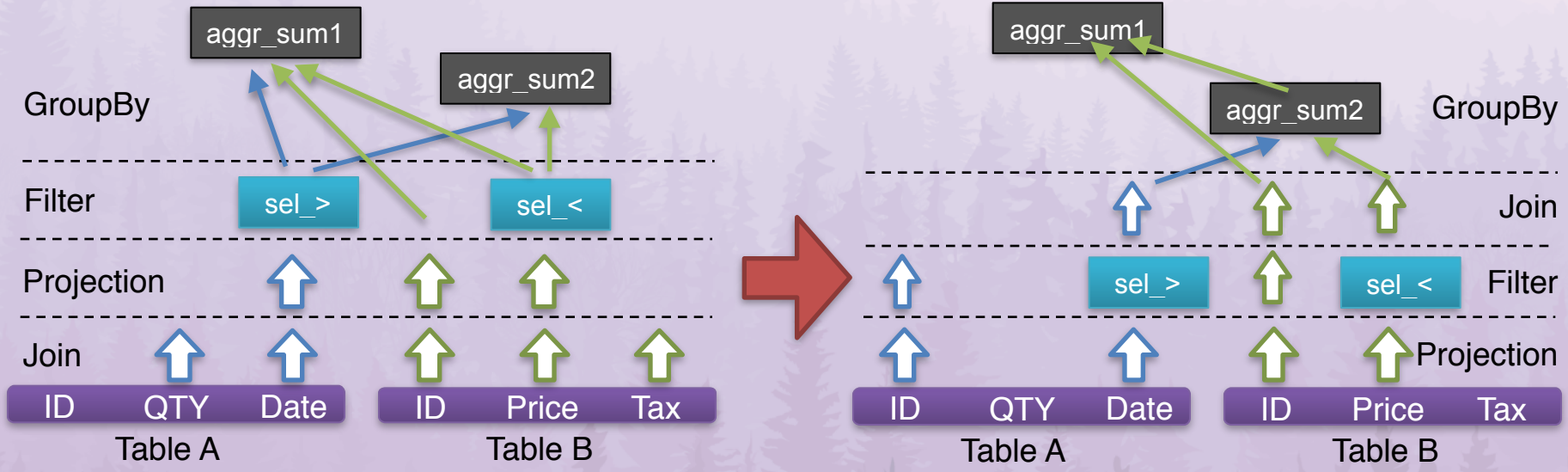
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo logical optimizer

- Cost-based join ordering
- Projection/Filter push down & Duplicated expression removal



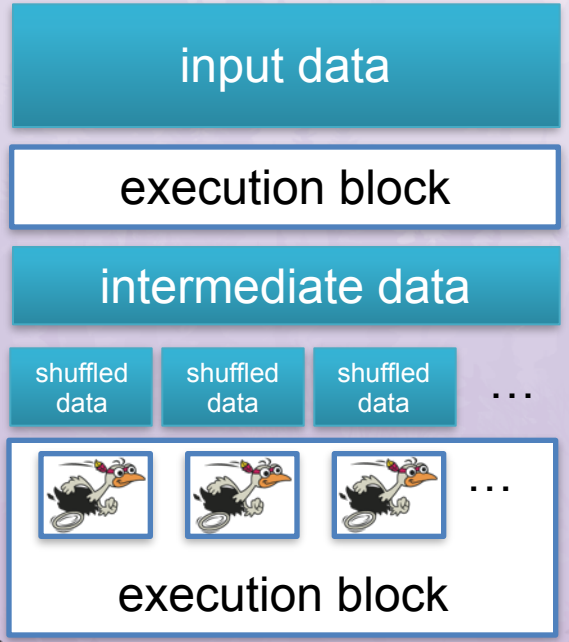
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo progressive optimization

- dynamically adjust number of tasks



unknown priorly

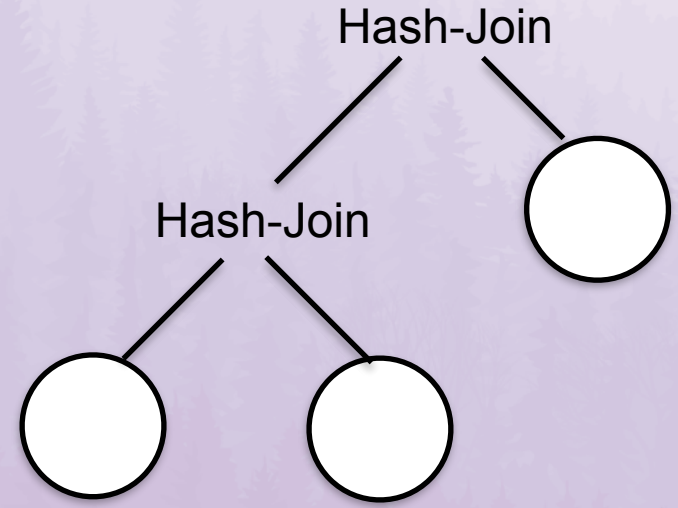
how many tasks (and workers)?

- estimate data size at planning time
- check size and adjust plan at execution time
- shuffle intermediate data over workers uniformly

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo progressive optimization

- dynamically adjust join order or type



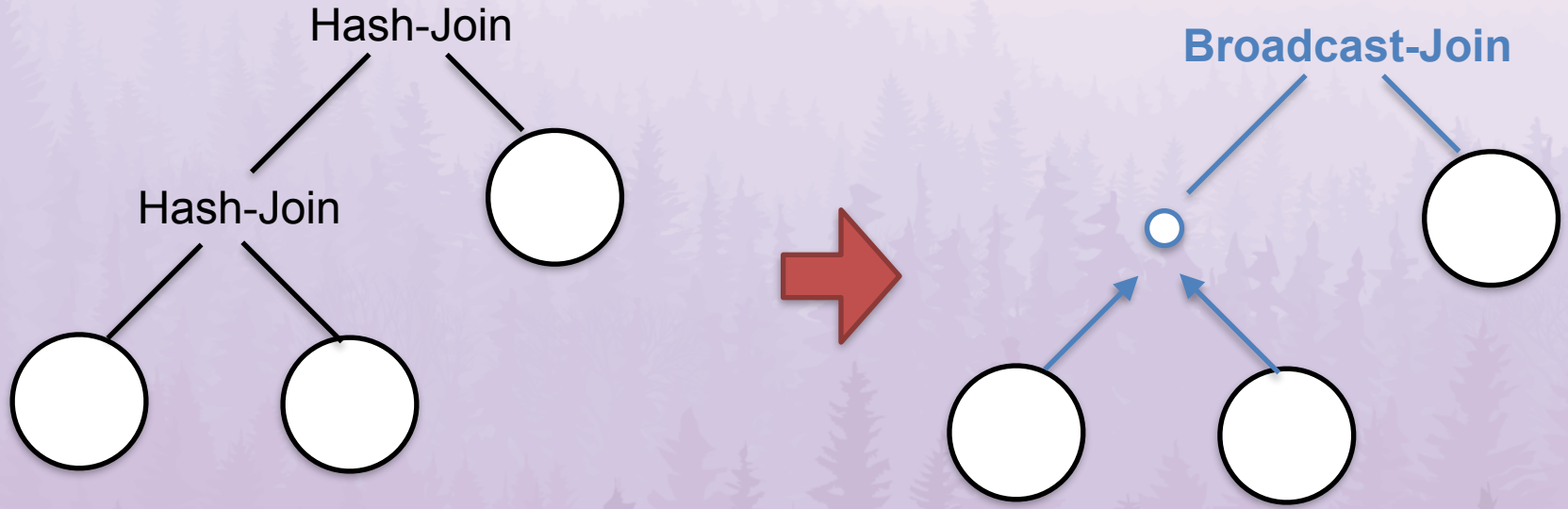
APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo progressive optimization

- dynamically adjust join order or type



APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo - what is improved past 9 months ?

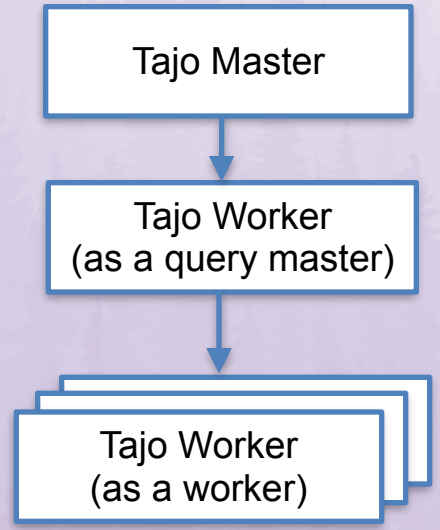
- Resource Manager
- Scheduler & Storage Manager
- Data types & Functions
- SQL Interface
- Management

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Tajo resource manager

- Fine resource allocation



TAJO-127
without YARN

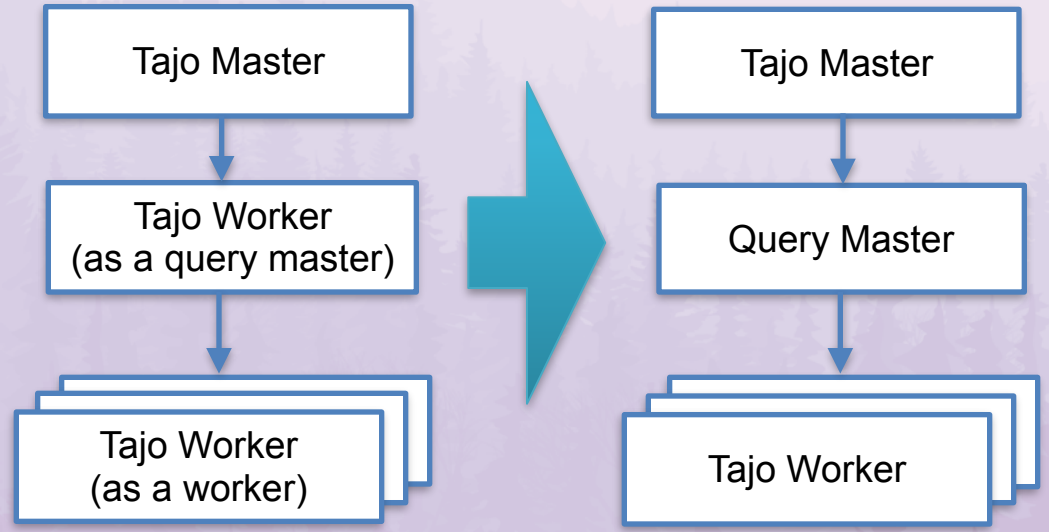
APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Tajo resource manager

- Fine resource allocation



TAJO-127
without YARN

TAJO-275
separating Query master

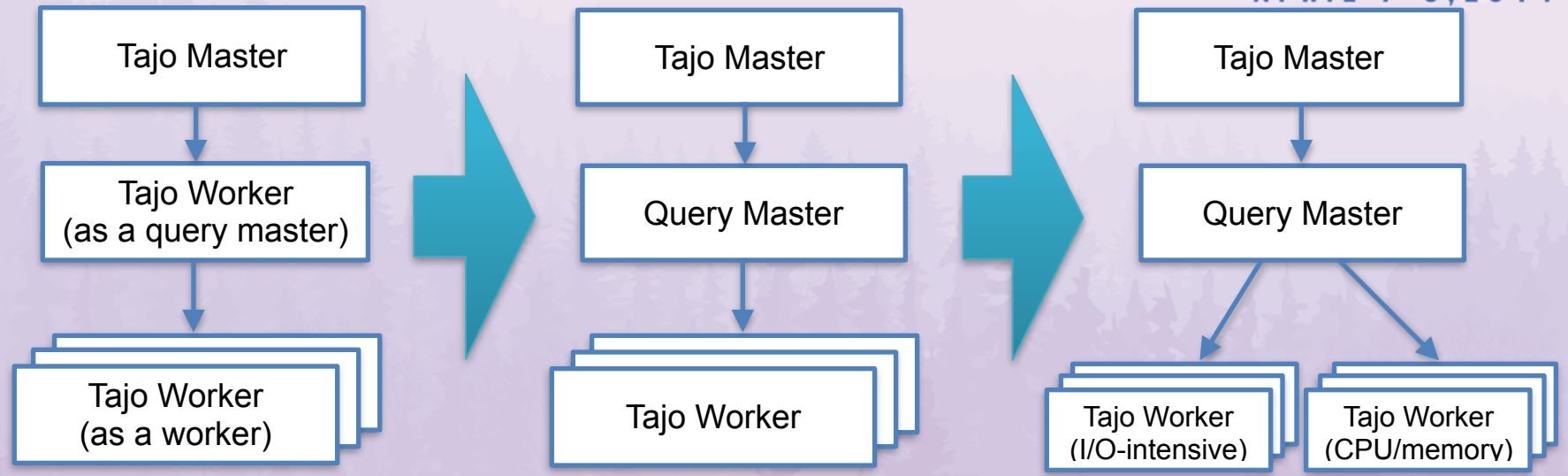
TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Tajo resource manager

- Fine resource allocation



TAJO-127
without YARN

TAJO-275
separating Query master

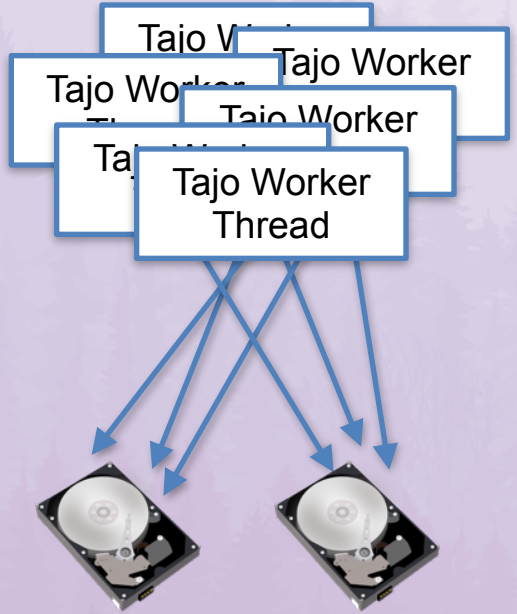
TAJO-317
elaborate resource allocation

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Scheduler & Storage manager

- disk-aware scheduling (volume info from HDFS-3672)



APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

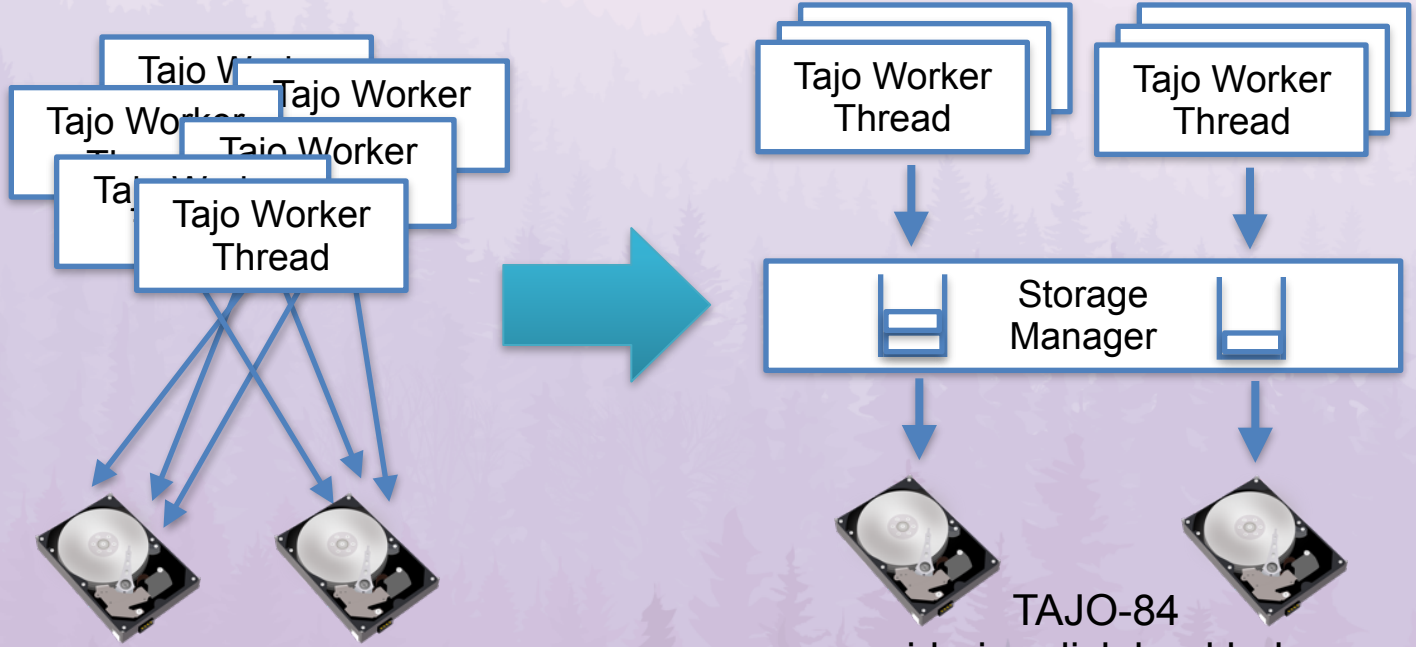
TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Scheduler & Storage manager

- disk-aware scheduling (volume info from HDFS-3672)



TAJO-84
considering disk load balance
TAJO-178
asynchronous scan

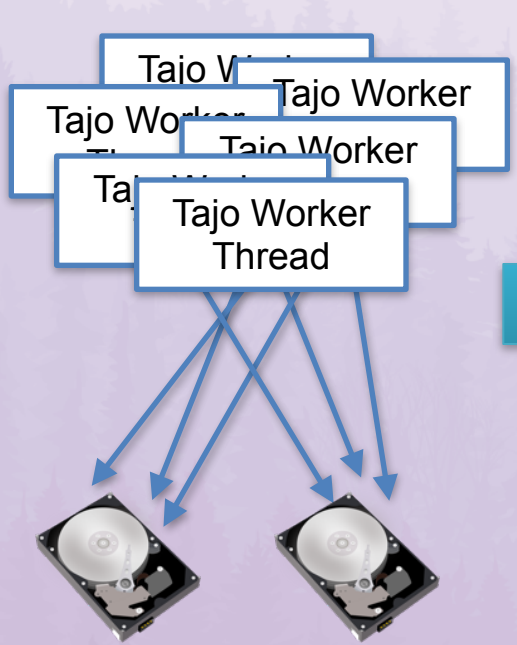
TAJO

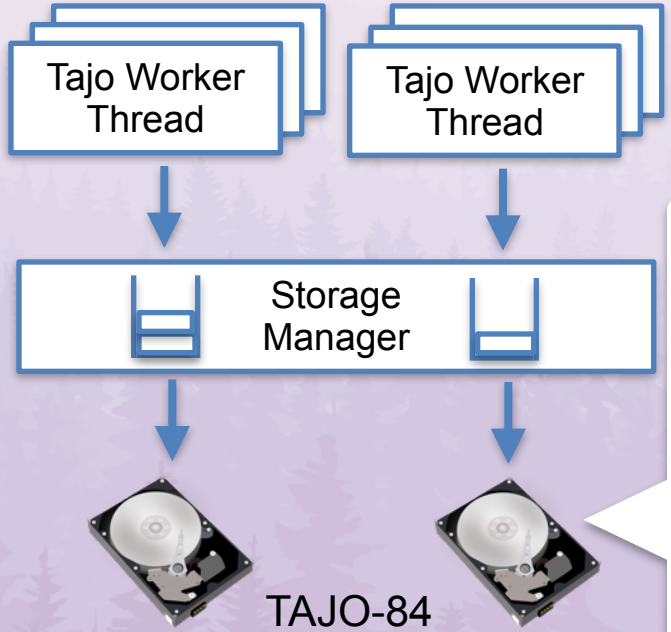
Presented For The Apache Foundation By
LINUX FOUNDATION

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Scheduler & Storage manager

- disk-aware scheduling (volume info from HDFS-3672)





TAJO-84
considering disk load balance
TAJO-178
asynchronous scan

TAJO-134
text compression
(gzip, snappy, lz4, bzip2)

TAJO-200
RCFile

TAJO-30
Parquet

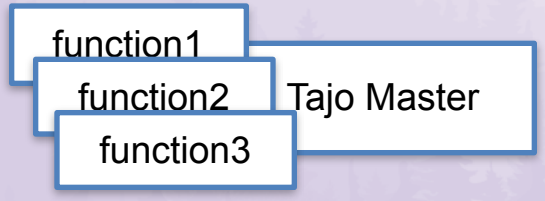
TAJO-435
intermediate file

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Functions & data types

- supporting more functions and UDFs



registered at startup
(class name is coded in source)

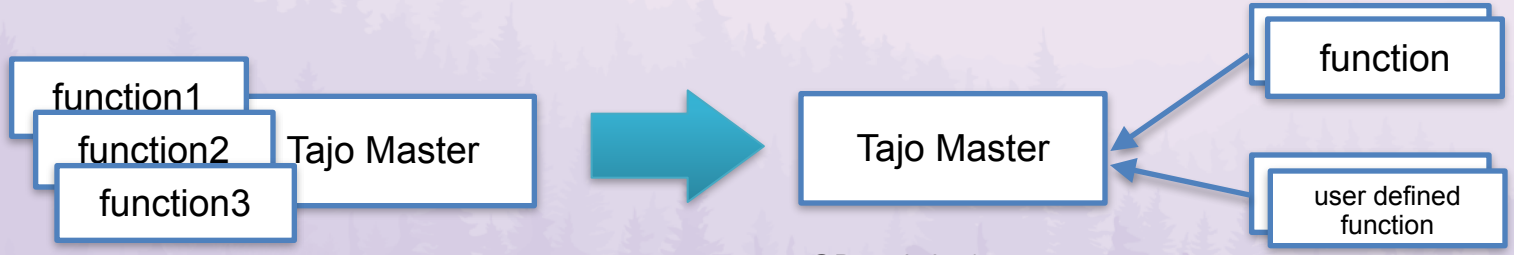
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Functions & data types

- supporting more functions and UDFs



registered at startup
(class name is coded in source)

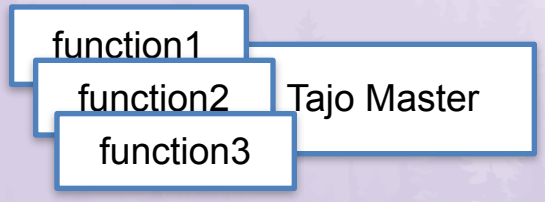
```
@Description(
  functionName = "to_timestamp",
  description = "Convert UNIX epoch to time stamp",
  example = "> SELECT to_timestamp(1389071574);\n"
        + "2014-01-07 14:12:54",
  returnType = TajoDataTypes.Type.TIMESTAMP,
  paramTypes = {@ParamTypes(paramTypes = {TajoDataTypes.Type.INT4}),
     @ParamTypes(paramTypes = {TajoDataTypes.Type.INT8})}
)
```

TAJO-408
Improve function system

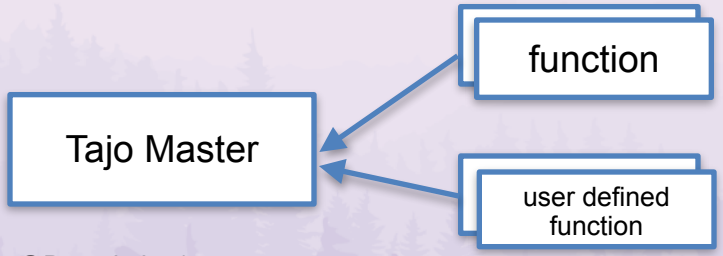
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Functions & data types

- supporting more functions and UDFs



registered at startup
(class name is coded in source)



automatic registration

runtime registration

```
@Description(
  functionName = "to_timestamp",
  description = "Convert UNIX epoch to time stamp",
  example = "> SELECT to_timestamp(1389071574);\n"
        + "2014-01-07 14:12:54",
  returnType = TajoDataTypes.Type.TIMESTAMP,
  paramTypes = {@ParamTypes(paramTypes = {TajoDataTypes.Type.INT4}),
      @ParamTypes(paramTypes = {TajoDataTypes.Type.INT8})}
)
```

description

TAJO-408
Improve function system

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

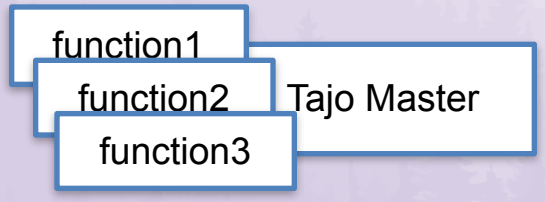
TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Functions & data types

- supporting more functions and UDFs



registered at startup
(class name is coded in source)

automatic registration

runtime registration

```
@Description(
  functionName = "to_timestamp",
  description = "Convert UNIX epoch to time stamp",
  example = "> SELECT to_timestamp(1389071574);\n"
        + "2014-01-07 14:12:54",
  returnType = TajoDataTypes.Type.TIMESTAMP,
  paramTypes = {@ParamTypes(paramTypes = {TajoDataTypes.Type.INT4}),
     @ParamTypes(paramTypes = {TajoDataTypes.Type.INT8})}
)
```

description

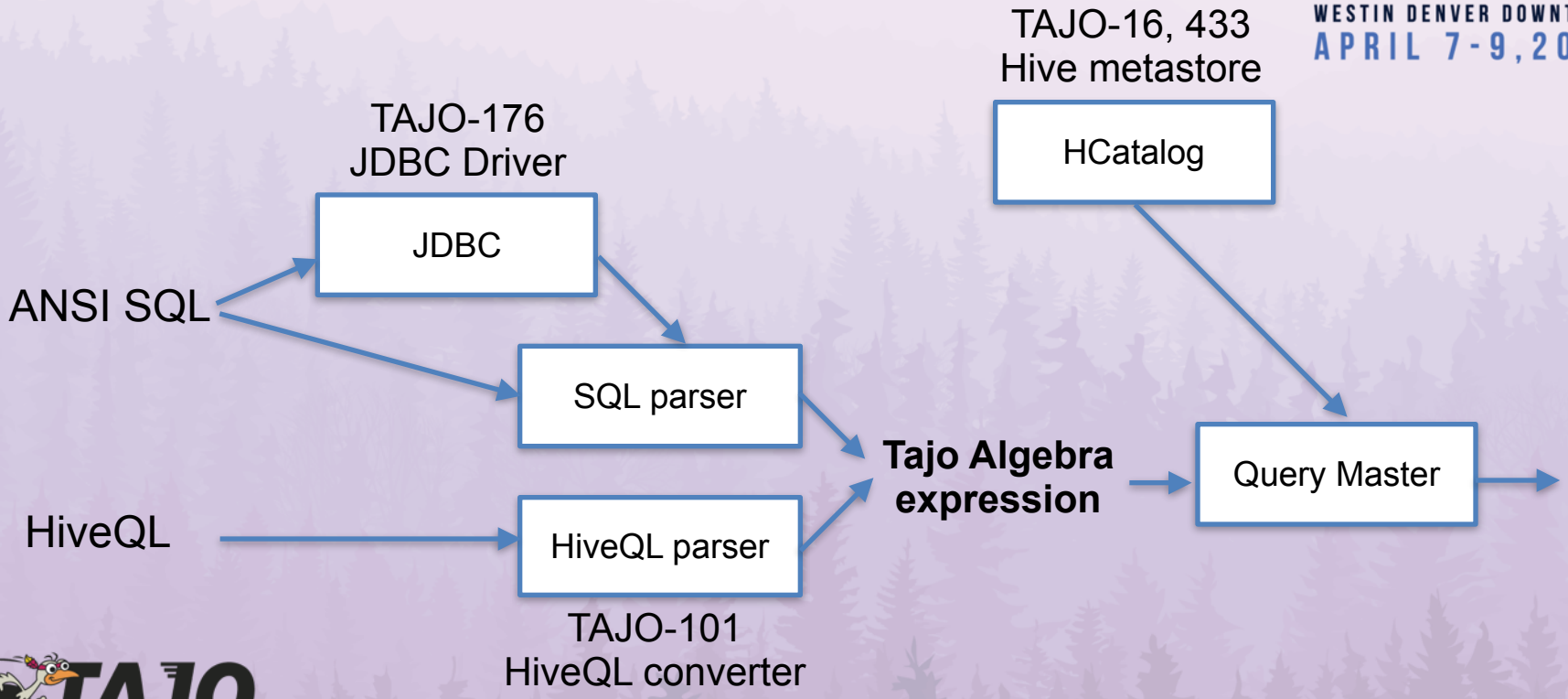
TAJO-408
Improve function system

TAJO-52
standard SQL
data types

TAJO

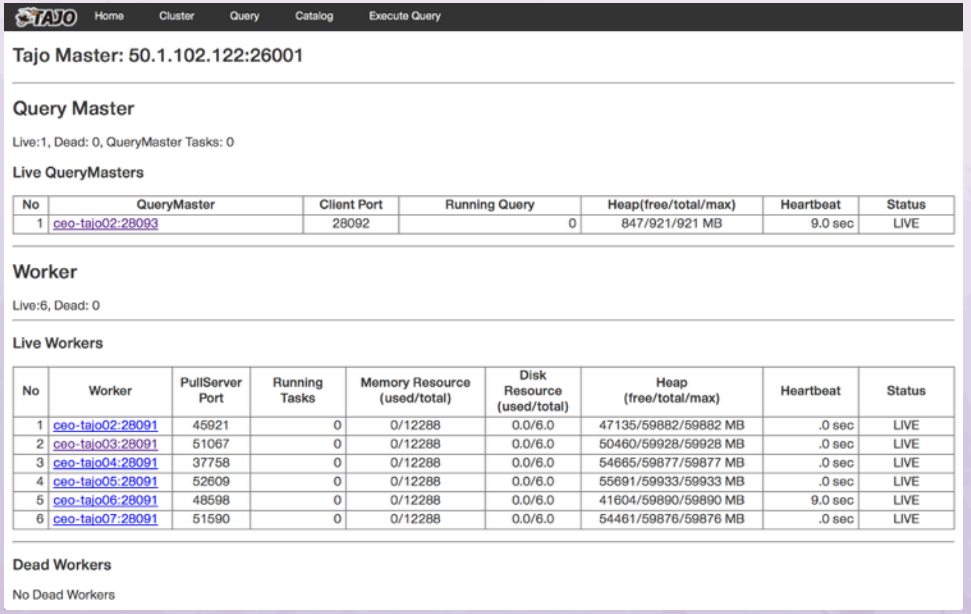
Presented For The Apache Foundation By
LINUX FOUNDATION

JDBC Driver, HCatalog



APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

Management



TAJO Home Cluster Query Catalog Execute Query

Tajo Master: 50.1.102.122:26001

Query Master

Live:1, Dead: 0, QueryMaster Tasks: 0

Live QueryMasters

No	QueryMaster	Client Port	Running Query	Heap(free/total/max)	Heartbeat	Status
1	ceo-tajo02:28093	28092	0	847/921/921 MB	9.0 sec	LIVE

Worker

Live:6, Dead: 0

Live Workers

No	Worker	PullServer Port	Running Tasks	Memory Resource (used/total)	Disk Resource (used/total)	Heap (free/total/max)	Heartbeat	Status
1	ceo-tajo02:28091	45921	0	0/12288	0.0/6.0	47135/59882/59882 MB	.0 sec	LIVE
2	ceo-tajo03:28091	51067	0	0/12288	0.0/6.0	50460/59928/59928 MB	.0 sec	LIVE
3	ceo-tajo04:28091	37758	0	0/12288	0.0/6.0	54665/59877/59877 MB	.0 sec	LIVE
4	ceo-tajo05:28091	52609	0	0/12288	0.0/6.0	55691/59933/59933 MB	.0 sec	LIVE
5	ceo-tajo06:28091	48598	0	0/12288	0.0/6.0	41604/59890/59890 MB	9.0 sec	LIVE
6	ceo-tajo07:28091	51590	0	0/12288	0.0/6.0	54461/59876/59876 MB	.0 sec	LIVE

Dead Workers

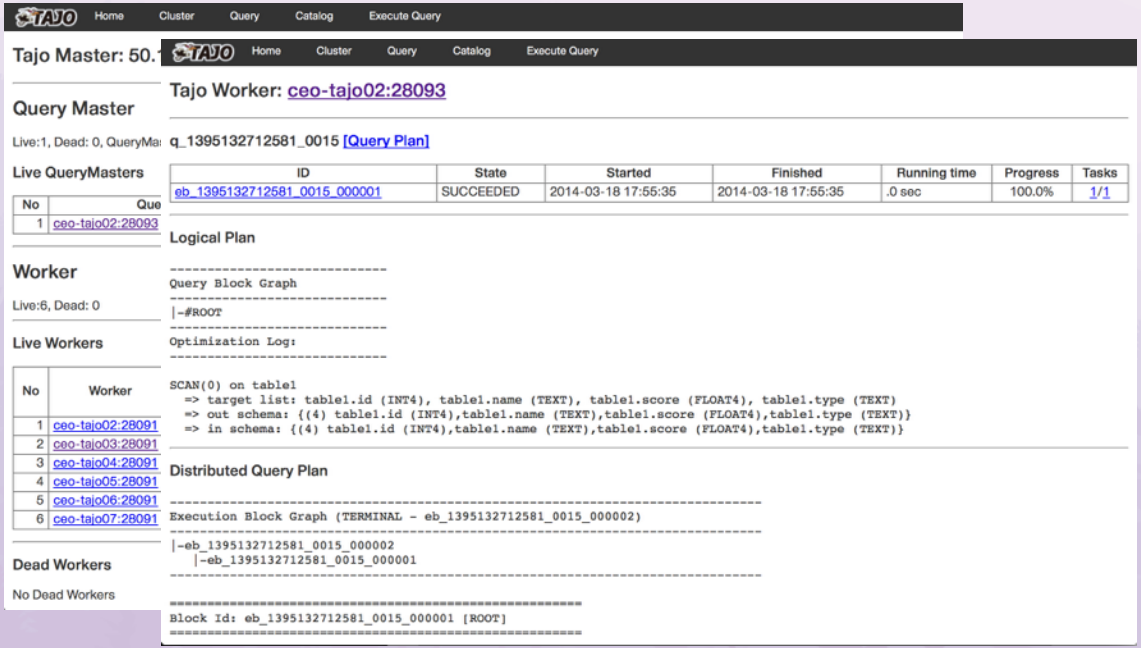
No Dead Workers

TAJO-239
Improving Web UI

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Management

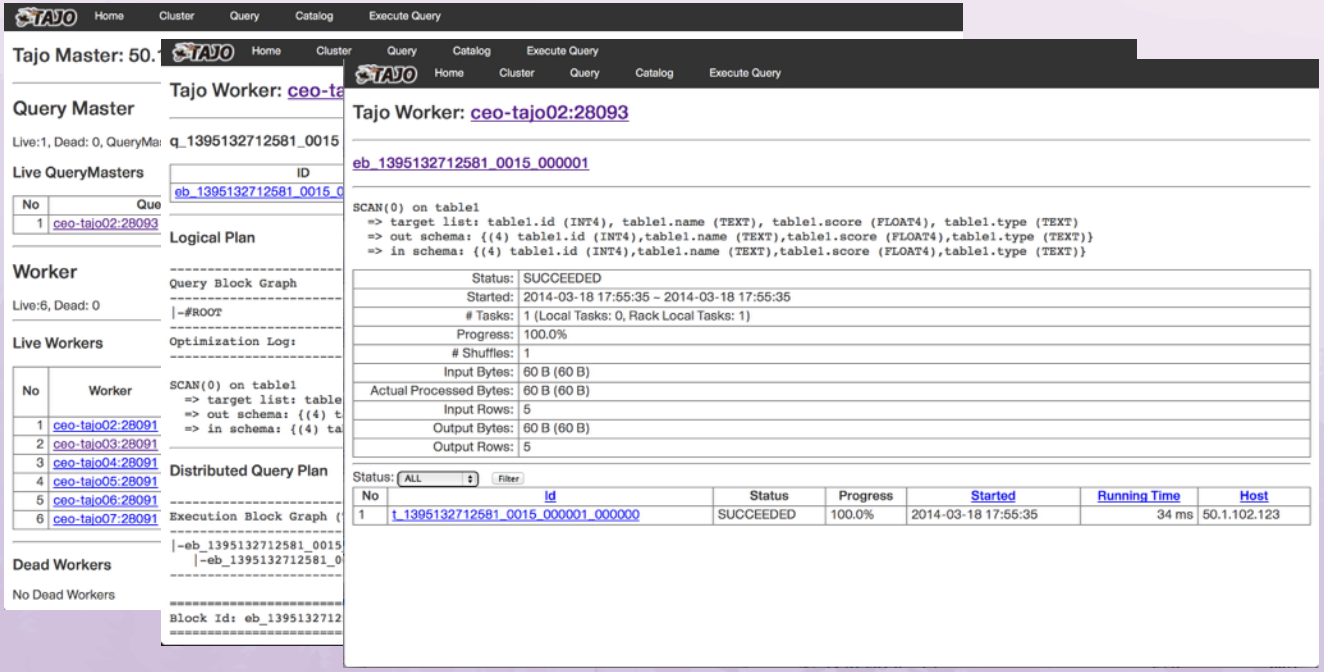


TAJO-564
Execution block progress

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Management



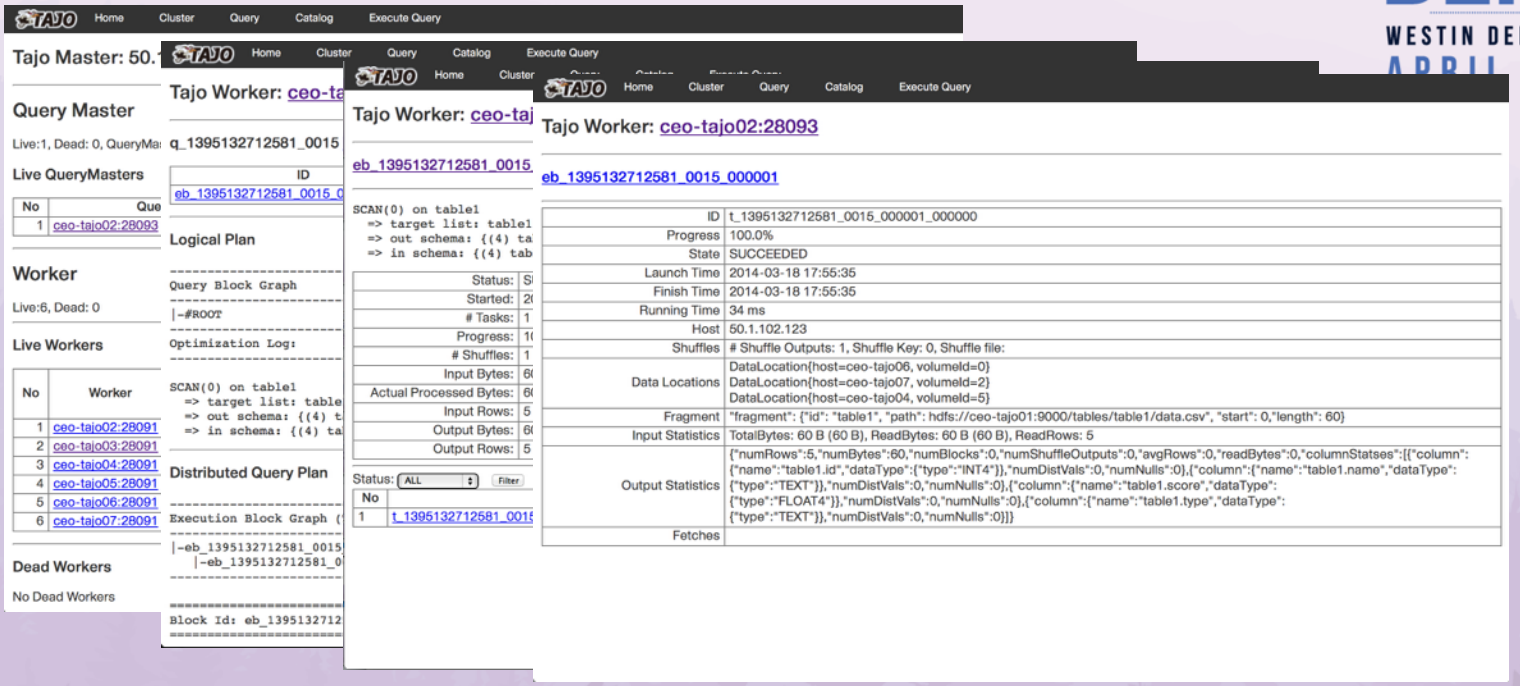
TAJO-589
Task progress

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

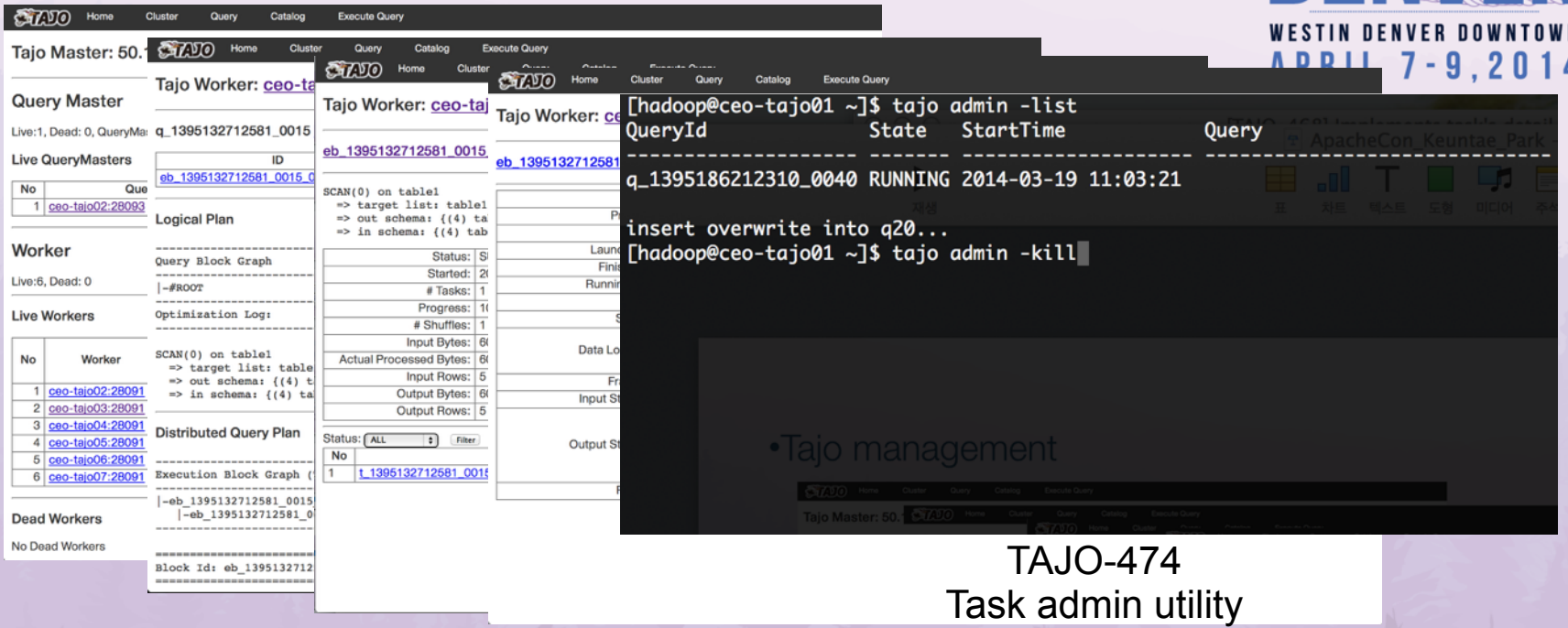
Management



TAJO-468
Task detail info

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
Presented For The Apache Foundation By
LINUX FOUNDATION

Management



TAJO-474
Task admin utility

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014
TAJO
Presented For The Apache Foundation By
LINUX FOUNDATION

And lots of Performance enhancement

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO-725 Broadcast JOIN should supports multiple tables

TAJO-717 Improve file splitting for large number of splits

TAJO-601 Improve distinct aggregation query processing

TAJO-584 Improve distributed merge sort

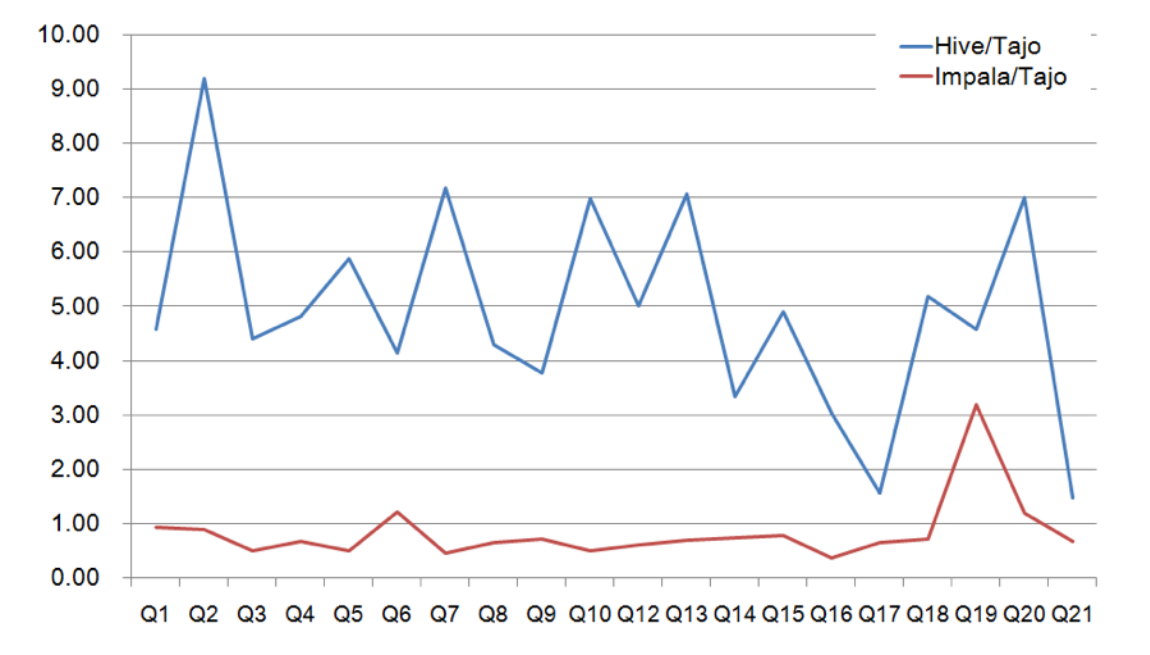
TAJO-36 Improve ExternalSortExec with N-merge sort and final pass omission

TAJO-345 MergeScanner should support projectable storages

...

Performance

- TPC-H



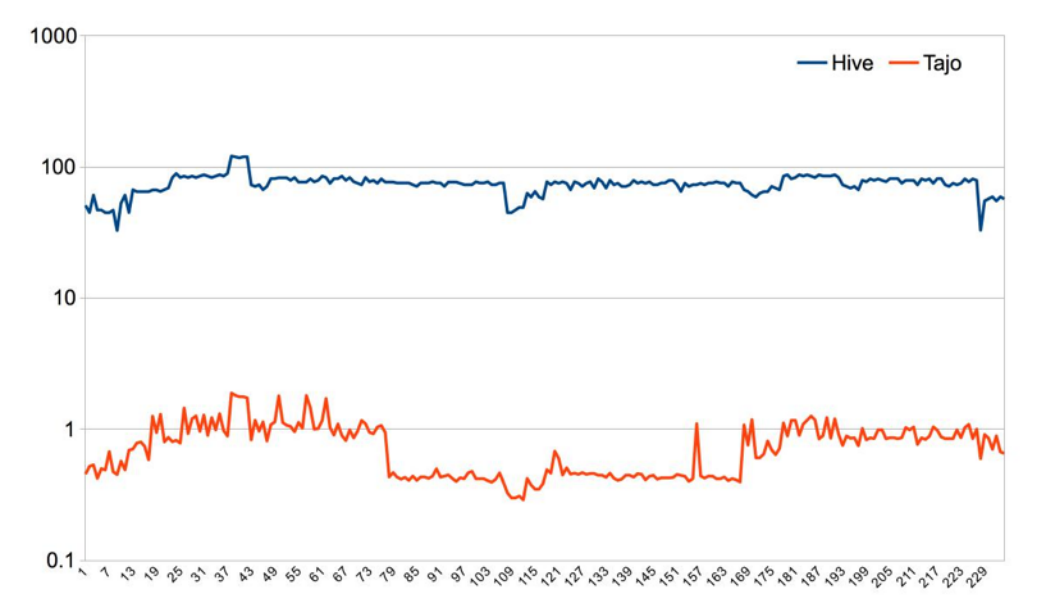
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Performance

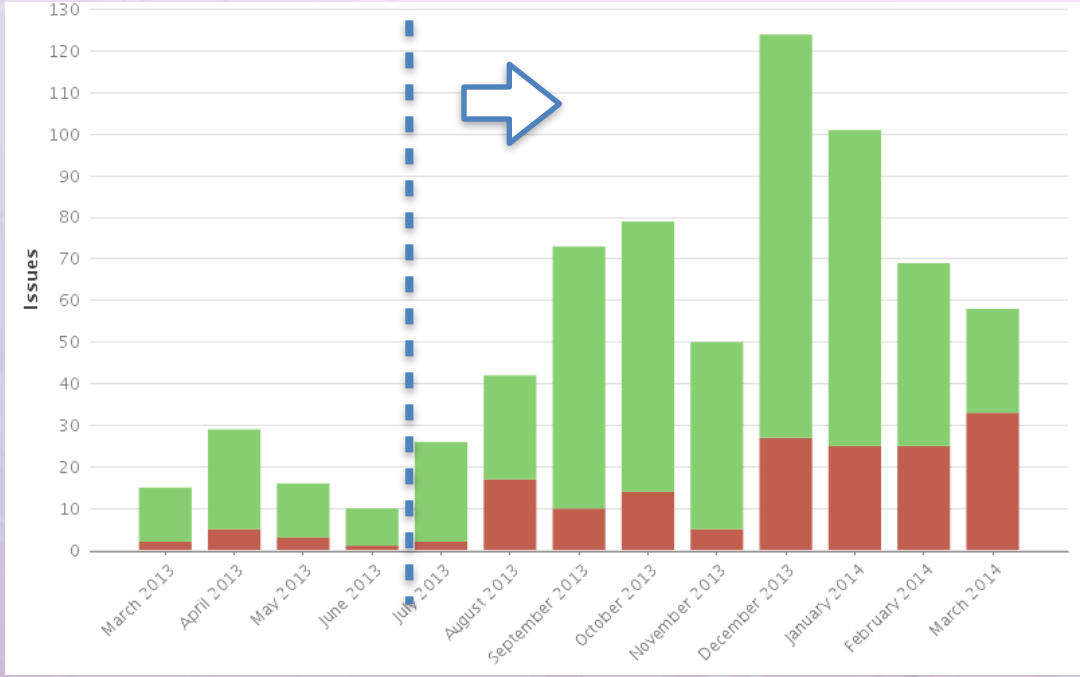
- OLAP reporting - relatively small data



TAJO

win-win between company and community

- Community boom up



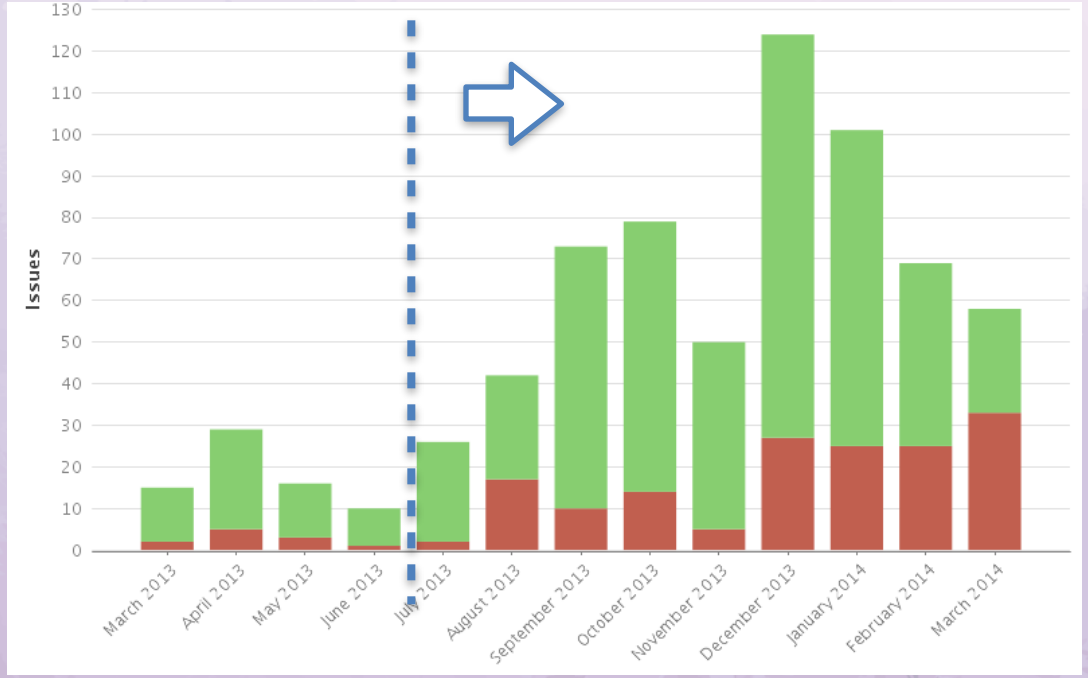
APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

win-win between company and community

- Community boom up



13 ⇨ 30

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

win-win between company and community

- Test in real working cluster
  - Mainly focusing on the scalability test & integration with existing IT systems
  - Finding bugs and function requirements, also

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

win-win between company and community

- Test in real working cluster
  - Mainly focusing on the scalability test & integration with existing IT systems
  - Finding bugs and function requirements, also

TAJO-691 HashJoin or HashAggregation is too slow if there is many unique keys
TAJO-675 maximum frame size of frameDecoder should be increased
TAJO-673 Assign proper number of tasks when inserting into partitioned table
TAJO-650 Repartitioner::scheduleHashShuffledFetches should adjust the number of tasks
TAJO-647 Work unbalance on disk scheduling of DefaultScheduler
TAJO-292 Too many intermediate partition files
TAJO-283 Add table partitioning
TAJO-592 HCatalogStore should supports RCFile and default hive field delimiter.
…

win-win between company and community

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

win-win between company and community

- efficient development and operation
- human networking
- brand value up - recruiting

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

TAJO

Presented For The Apache Foundation By
LINUX FOUNDATION

Future Works

- Nested data model (parquet model)
- more SQL compatible
  - window functions, IN, EXIST
- Multi-tenancy
- push shuffle (no materialization)
  - use selectively between push and pull shuffle
    - push shuffle: performance
    - pull shuffle: resilience, schedulability

APACHECON DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

Presented For The Apache Foundation By
LINUX FOUNDATION

Q & A

APACHECON
DENVER
WESTIN DENVER DOWNTOWN
APRIL 7-9, 2014

- Getting Started
  - http://tajo.apache.org/tajo-0.2.0-doc.html#GettingStarted
- Checkout the development branch
  - http://tajo.apache.org/downloads.html
- Jira - Issue Tracker
  - https://issues.apache.org/jira/browse/TAJO
- Join the mailing list
  - dev@tajo.apache.org

Presented For The Apache Foundation By
LINUX FOUNDATION