Netflix: Petabyte Scale Analytics Infrastructure in the Cloud

Daniel C. Weeks
Tom Gianos

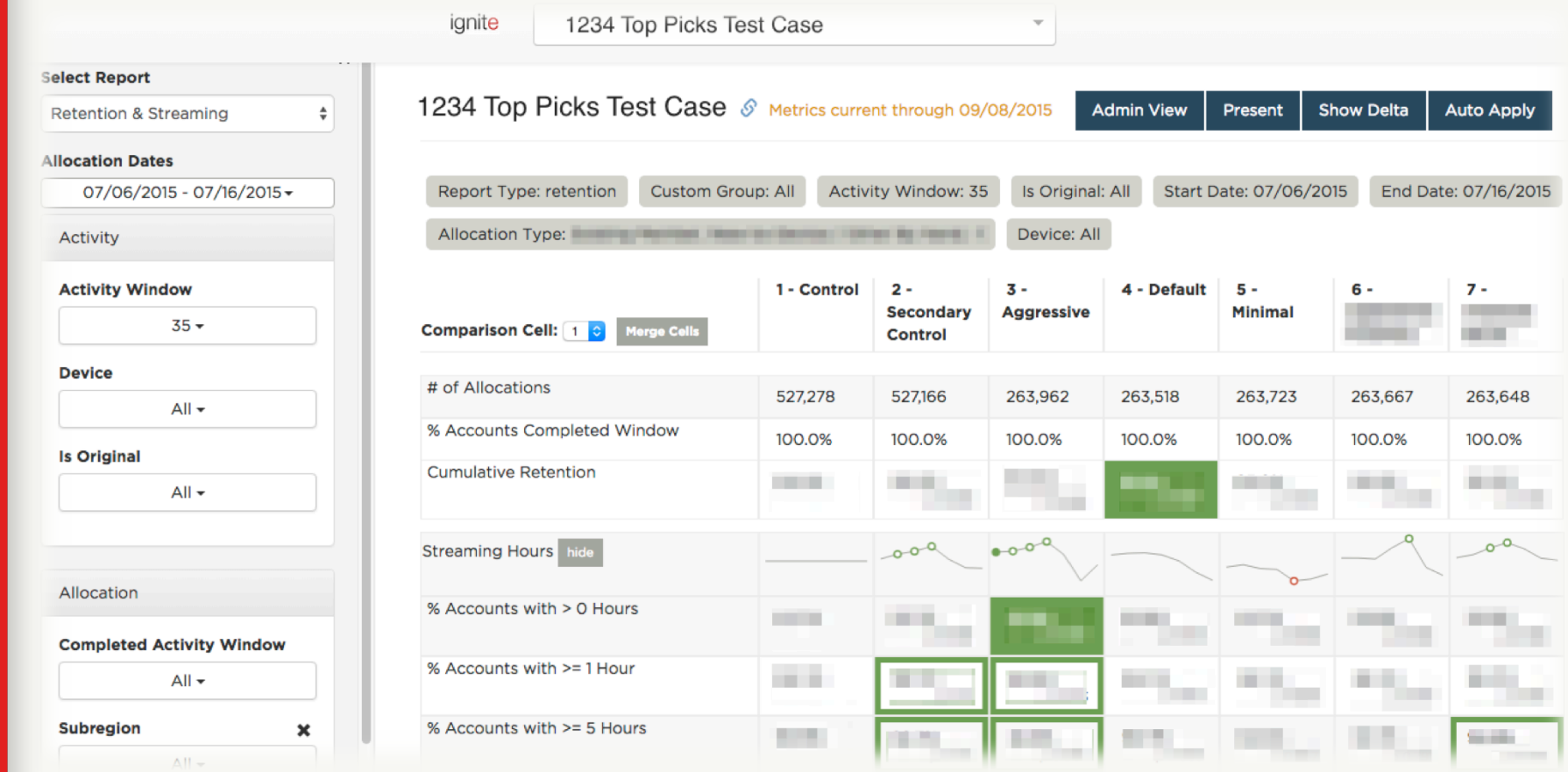
NETFLIX

Overview

- Data at Netflix
- Netflix Scale
- Platform Architecture
- Data Warehouse
- Genie
- Q&A

Data at Netflix


PLAN?
YES
NO
back up!
Save
W.C.C
proceed to back up
STEPS: (TO STORAGE)
GEN INFO:
Payback
RESULTS:
threshold
GAIN
RES. 0,17%
5.2%


ignite
1234 Top Picks Test Case
Select Report
Retention & Streaming
Allocation Dates
07/06/2015 - 07/16/2015
Activity
Activity Window
35
Device
All
Is Original
All
Allocation
Completed Activity Window
All
Subregion
1234 Top Picks Test Case
Metrics current through 09/08/2015
Admin View
Present
Show Delta
Auto Apply
Report Type: retention
Custom Group: All
Activity Window: 35
Is Original: All
Start Date: 07/06/2015
End Date: 07/16/2015
Allocation Type:
Device: All
Comparison Cell: 1
Merge Cells
1 - Control
2 - Secondary Control
3 - Aggressive
4 - Default
5 - Minimal
6 -
7 -
# of Allocations
527,278
527,166
263,962
263,518
263,723
263,667
263,648
% Accounts Completed Window
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
100.0%
Cumulative Retention
Streaming Hours
hide
% Accounts with > 0 Hours
% Accounts with >= 1 Hour
% Accounts with >= 5 Hours

UNITED STATES

The Netflix ISP Speed Index is a measure of **prime time Netflix performance on particular ISPs** (internet service providers) around the globe, and not a measure of overall performance for other services/data that may travel across the specific ISP network.

ISP LEADERBOARD - FEBRUARY 2016

SHOW SMALLER ISPS

RANK	ISP	SPEED Mbps		PREVIOUS Mbps	RANK CHANGE	TYPE (Fiber, Cable, DSL, Satellite, Wireless)
1	Verizon - FiOS	3.79		3.88		
2	Cox	3.71		3.85		
3	Bright House	3.69		3.53	+6	
4	Cablevision - Optimum	3.69		3.82	-1	
5	Comcast	3.65		3.72	-1	
6	Charter	3.60		3.65		
7	Mediacom	3.58		3.68	-2	

My List

NETFLIX ORIGINAL

CHEF'S TABLE

★★★★★ 2015 TV-14 1 Season

Resume

S1:E1 "Massimo Bottura"

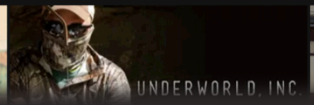
26 of 54m

By blending Italian tradition and artful modernity, chef Massimo Bottura's Osteria Francescana has been ranked the third best restaurant in the world.

MY LIST

OVERVIEW EPISODES TRAILERS MORE LIKE THIS DETAILS

TV Shows

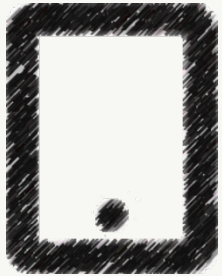


Our Biggest Challenge is Scale

Netflix Key Business Metrics

86+ million members

Global



1000+ devices supported

125+ million hours / day

Netflix Key Platform Metrics

500B Events

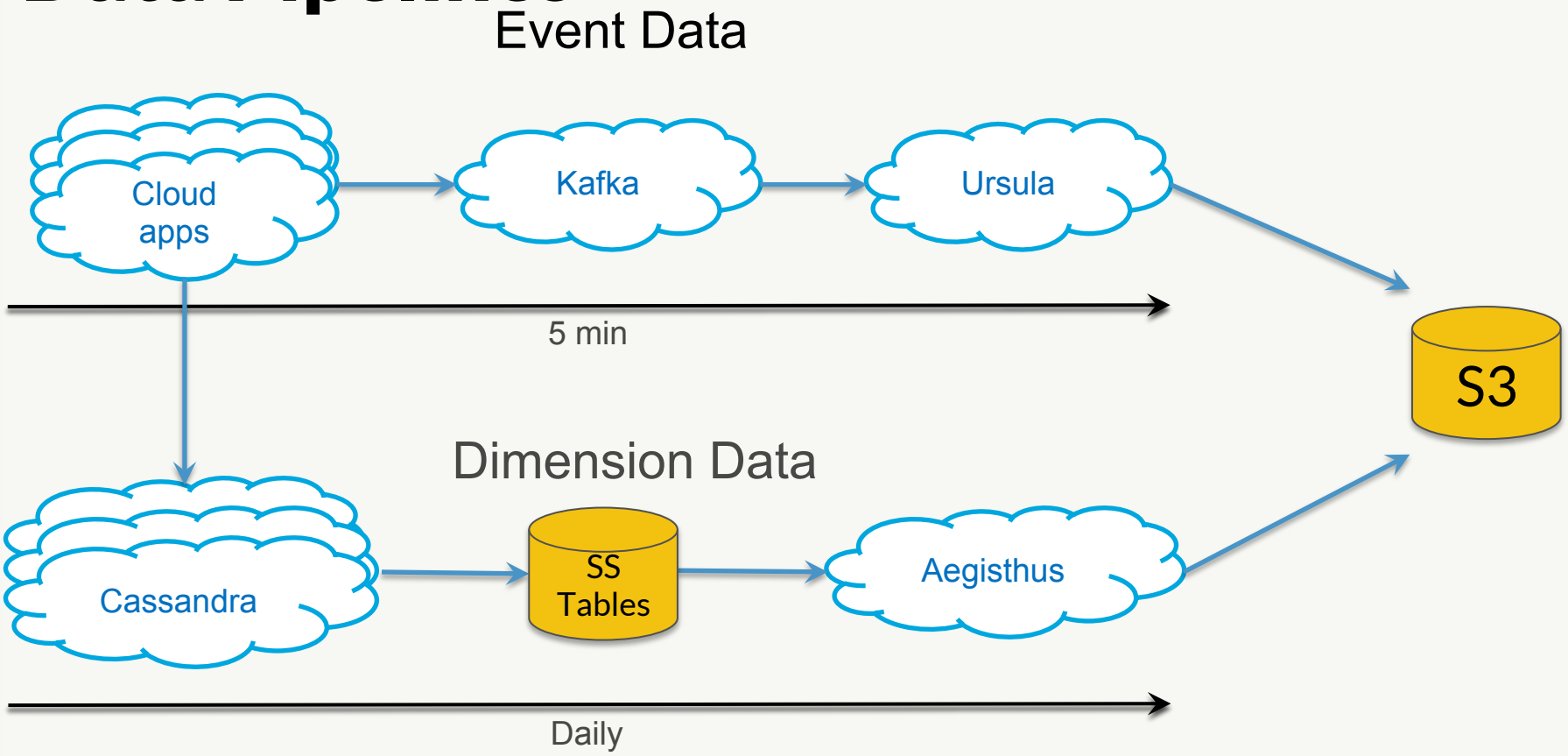
60 PB DW

Read 3PB

Write 500TB

Big Data Platform Architecture

Data Pipelines

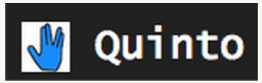


Interface

Big Data Portal

Big Data API

Tools



Transport | Visualization | Quality | Workflow Vis | Job/Cluster Vis

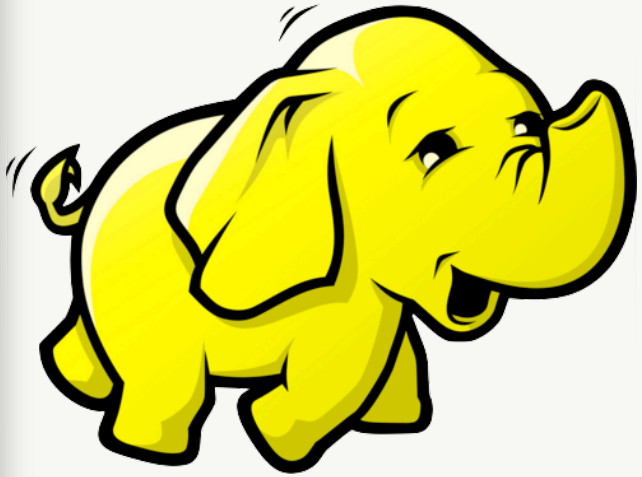
Service

Orchestration

Metacat Metadata

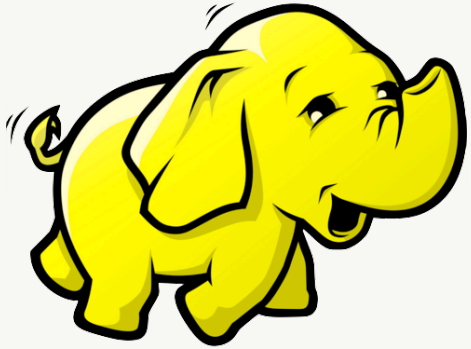
Compute

Storage



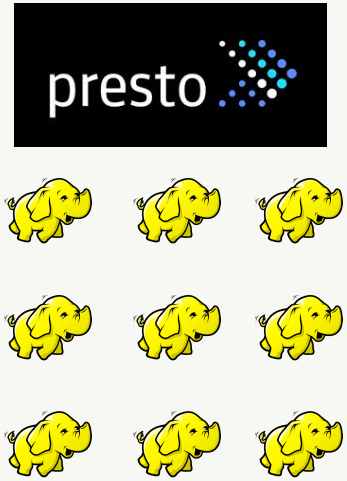
Production

~2300 d2.4xl



Ad-hoc

~1200 d2.4xl



Other

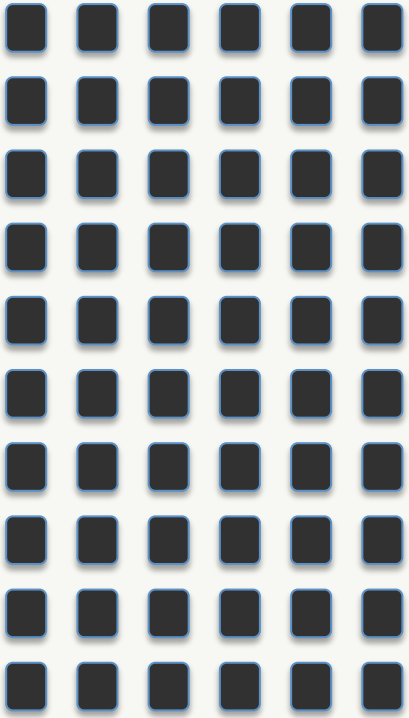
S3 Data Warehouse

Why S3?

- Lots of 9’s
- Features not available in HDFS
- Decouple Compute and Storage

Decoupled Scaling

Warehouse Size

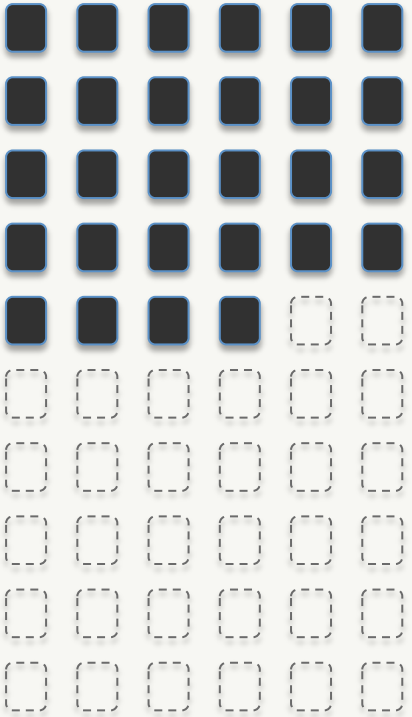


All Clusters

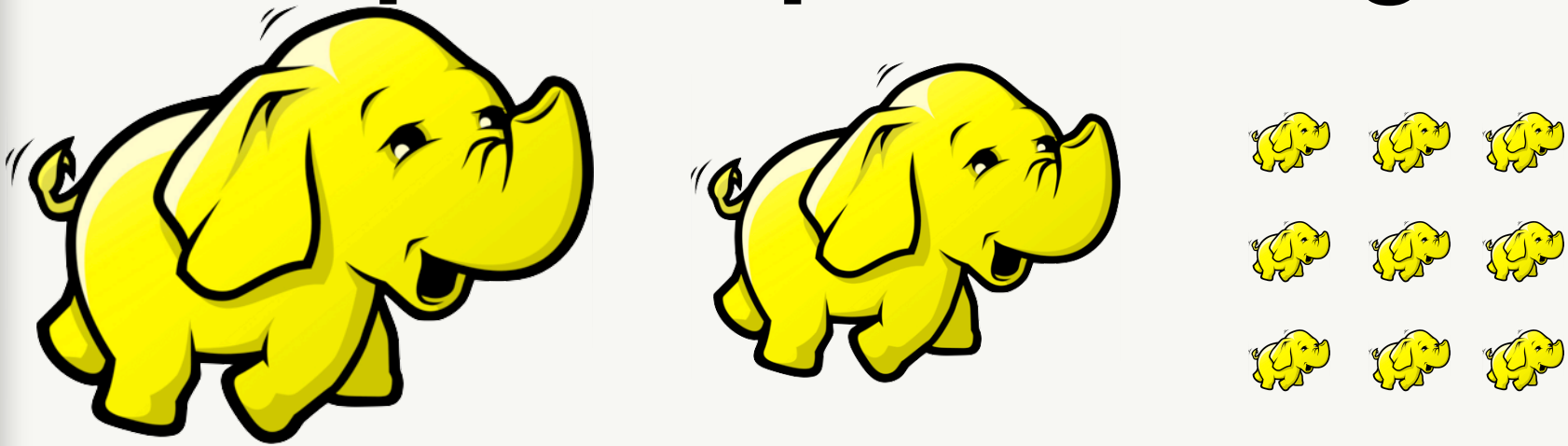
3x Replication

No Buffer

HDFS Capacity



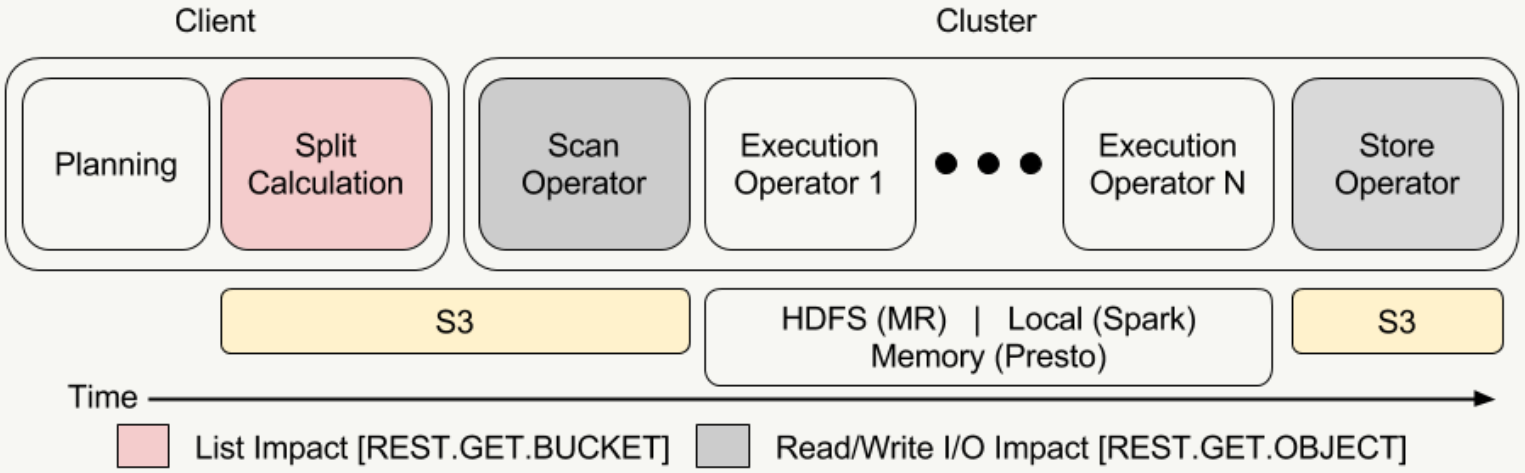
Decouple Compute / Storage



Tradeoffs - Performance

- Split Calculation (Latency)
  - Impacts job start time
  - Executes off cluster
- Table Scan (Latency + Throughput)
  - Parquet seeks add latency
  - Read overhead and available throughput
- Performance Converges with Volume and Complexity

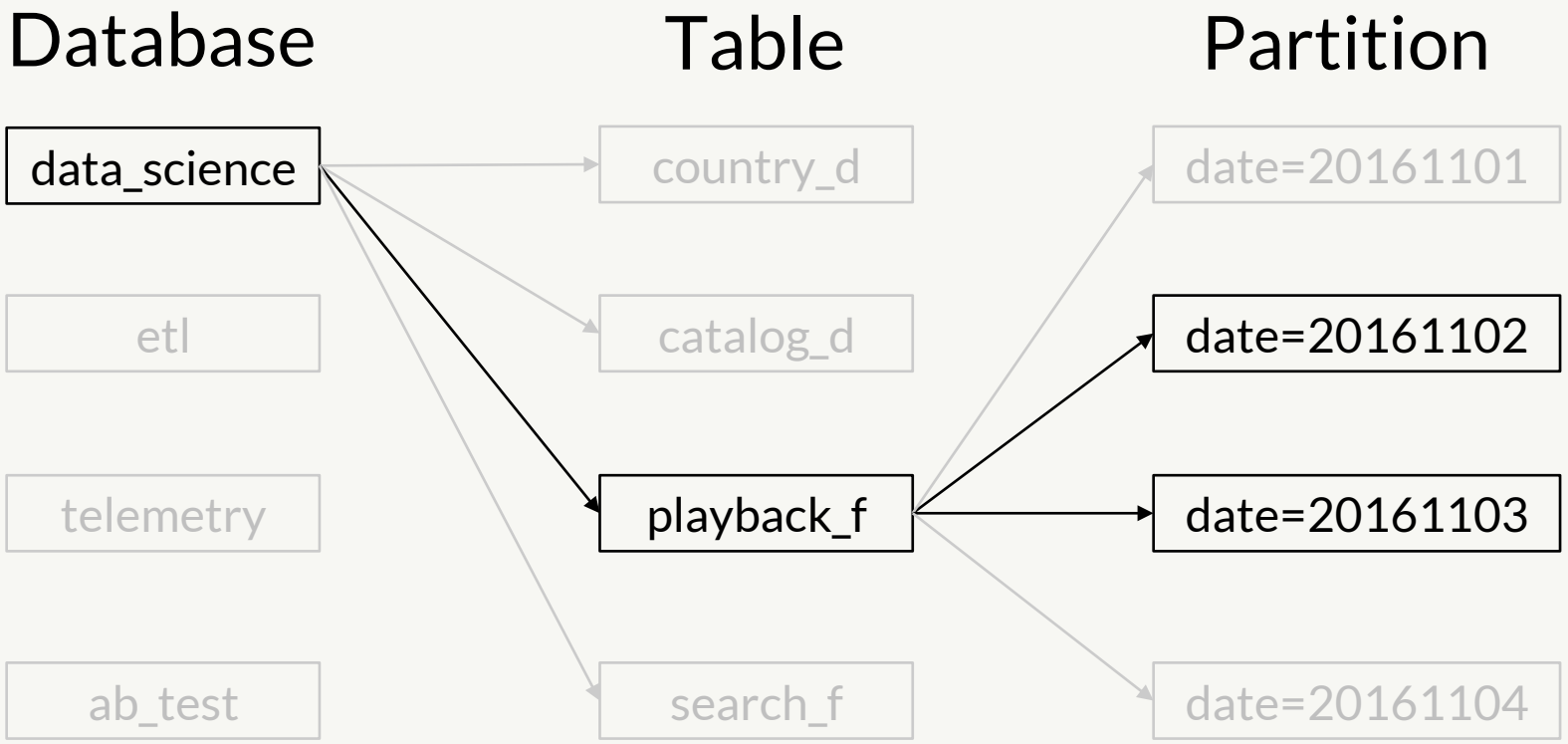
Tradeoffs - Performance



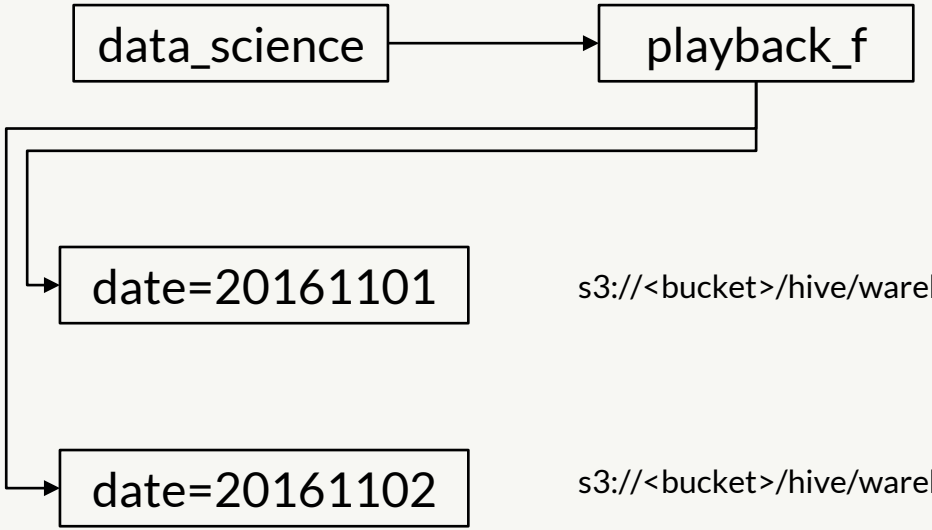
Metadata

- Metacat: Federated Metadata Service
- Hive Thrift Interface
- Logical Abstraction

Partitioning - Less is More



Partition Locations



s3://<bucket>/hive/warehouse/data_science.db/playback_f/dateint=20161101/…

s3://<bucket>/hive/warehouse/data_science.db/playback_f/dateint=20161102/…

Parquet

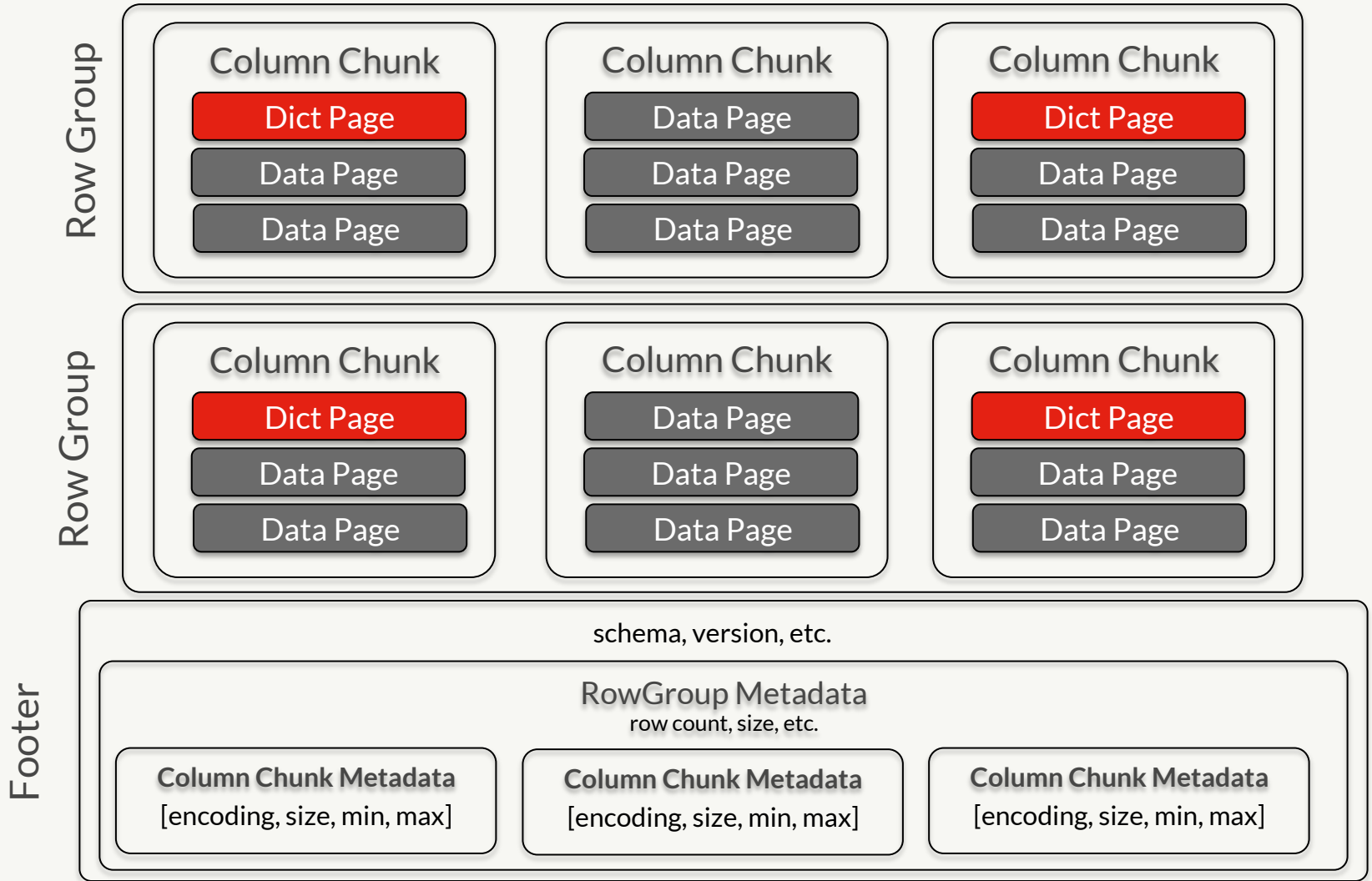
Parquet File Format

Column Oriented

- Store column data contiguously
- Improve compression
- Column projection

Strong Community Support

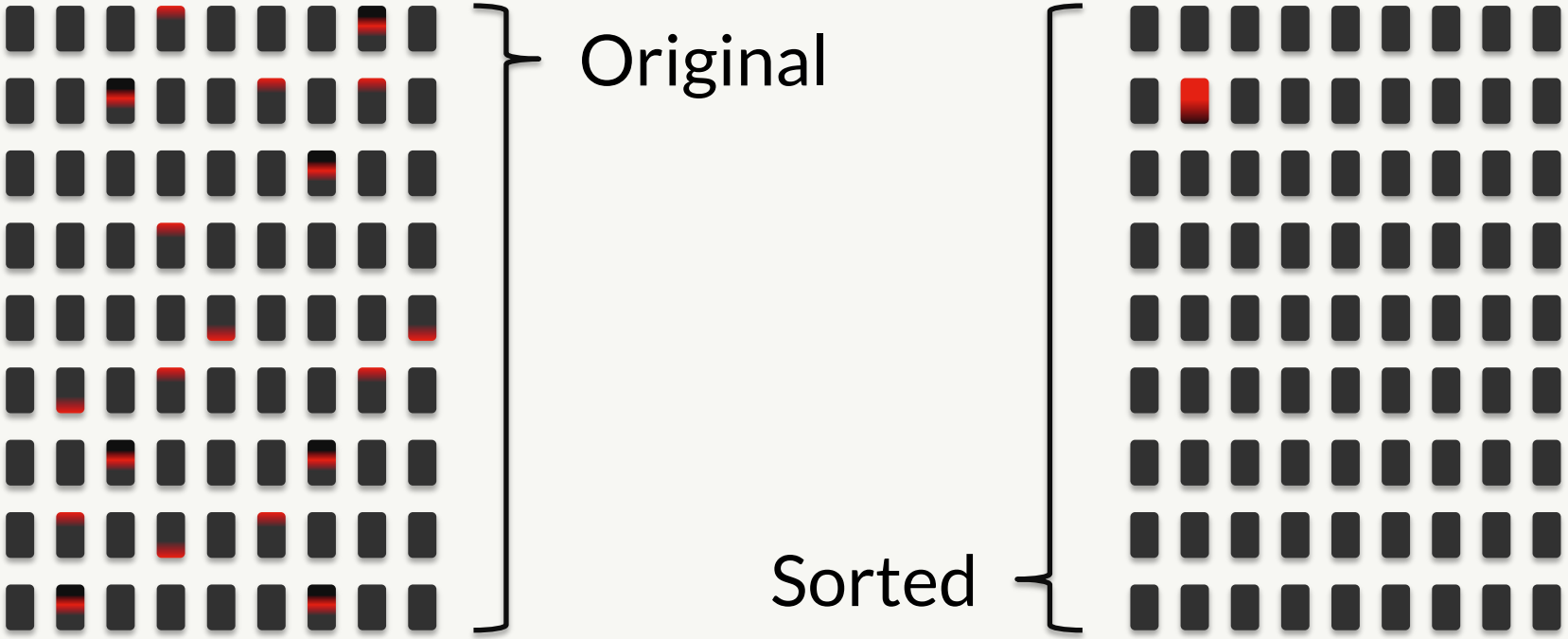
- Spark, Presto, Hive, Pig, Drill, Impala, etc.
- Works well with S3


Row Group
Column Chunk
Dict Page
Data Page
Data Page
Column Chunk
Data Page
Data Page
Data Page
Column Chunk
Dict Page
Data Page
Data Page
Row Group
Column Chunk
Dict Page
Data Page
Data Page
Column Chunk
Data Page
Data Page
Data Page
Column Chunk
Dict Page
Data Page
Data Page
Footer
schema, version, etc.
RowGroup Metadata
row count, size, etc.
Column Chunk Metadata
[encoding, size, min, max]
Column Chunk Metadata
[encoding, size, min, max]
Column Chunk Metadata
[encoding, size, min, max]

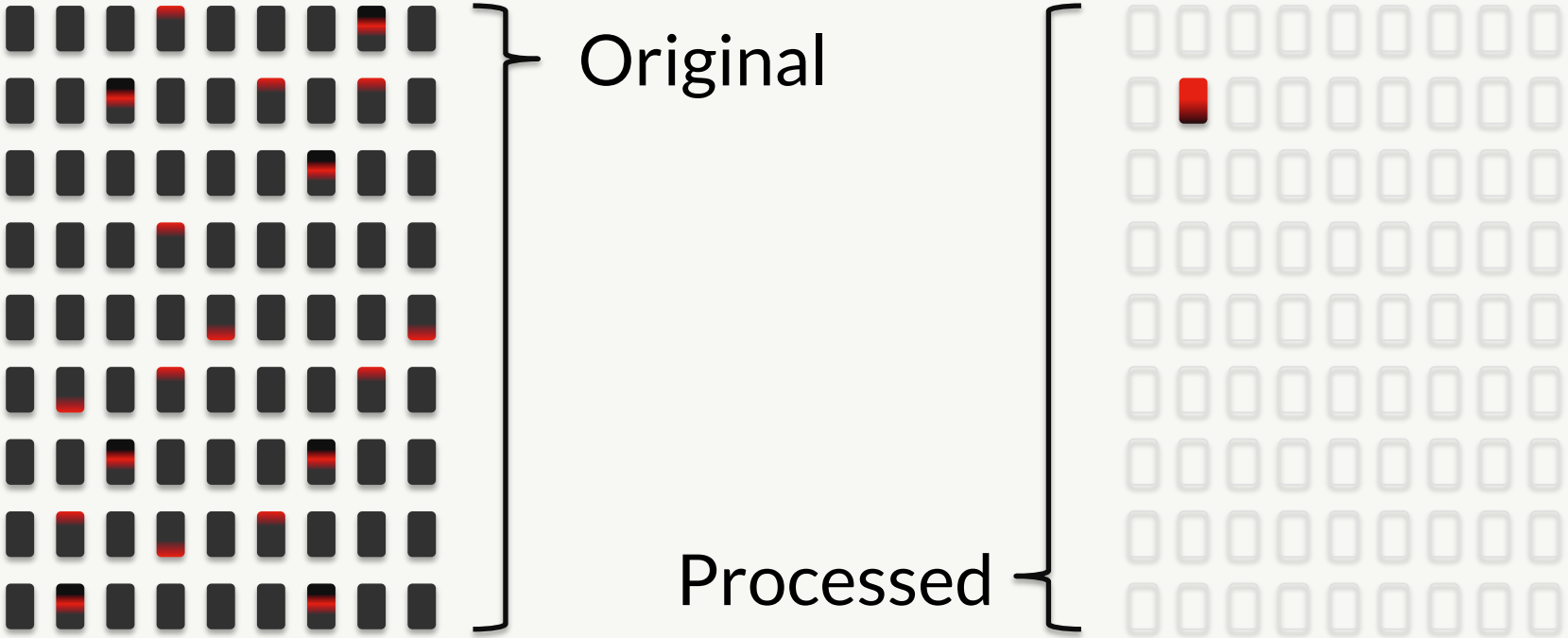
Staging Data

- Partition by low cardinality fields
- Sort by high cardinality predicate fields

Staging Data



Filtered

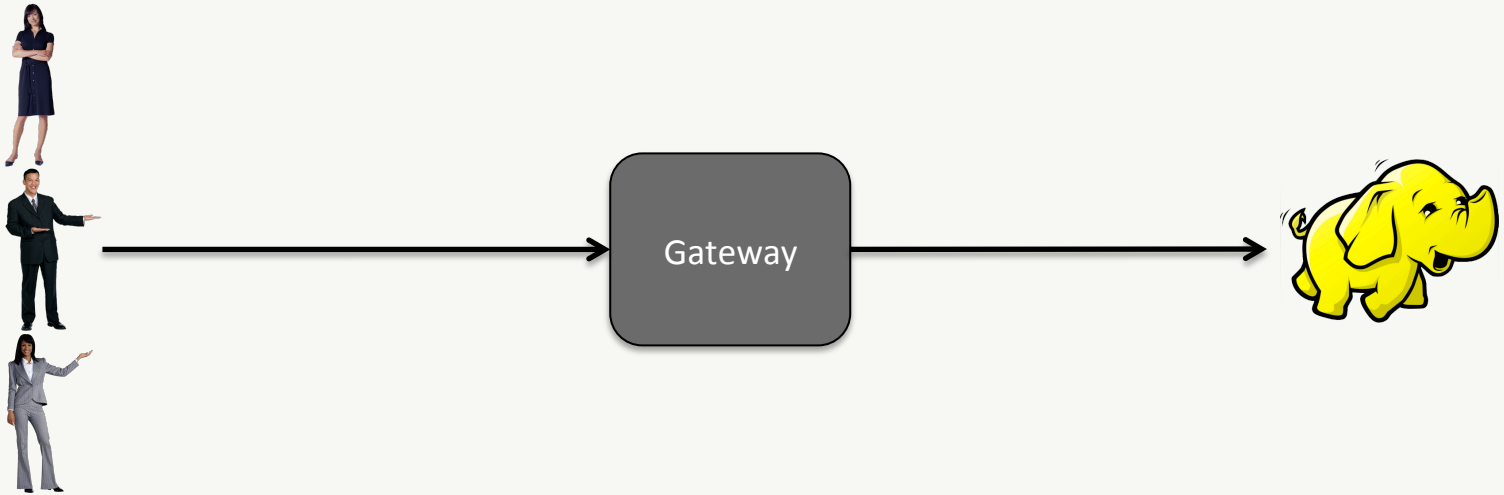


Parquet Tuning Guide

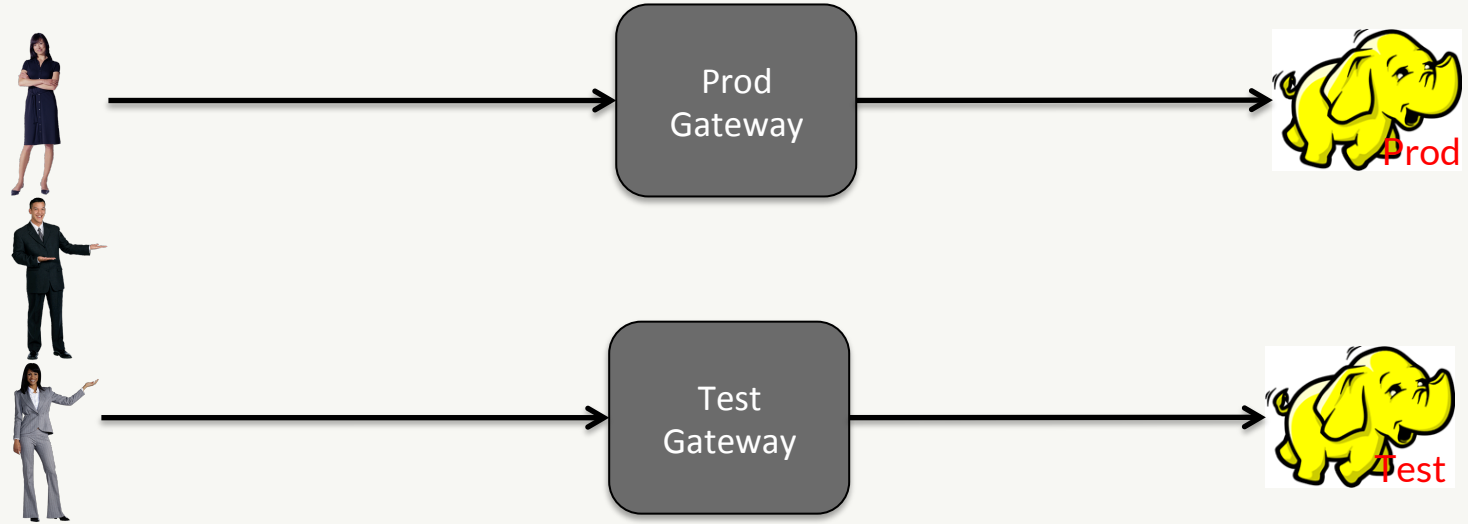
http://www.slideshare.net/RyanBlue3/parquet-performance-tuning-the-missing-guide


GENIE

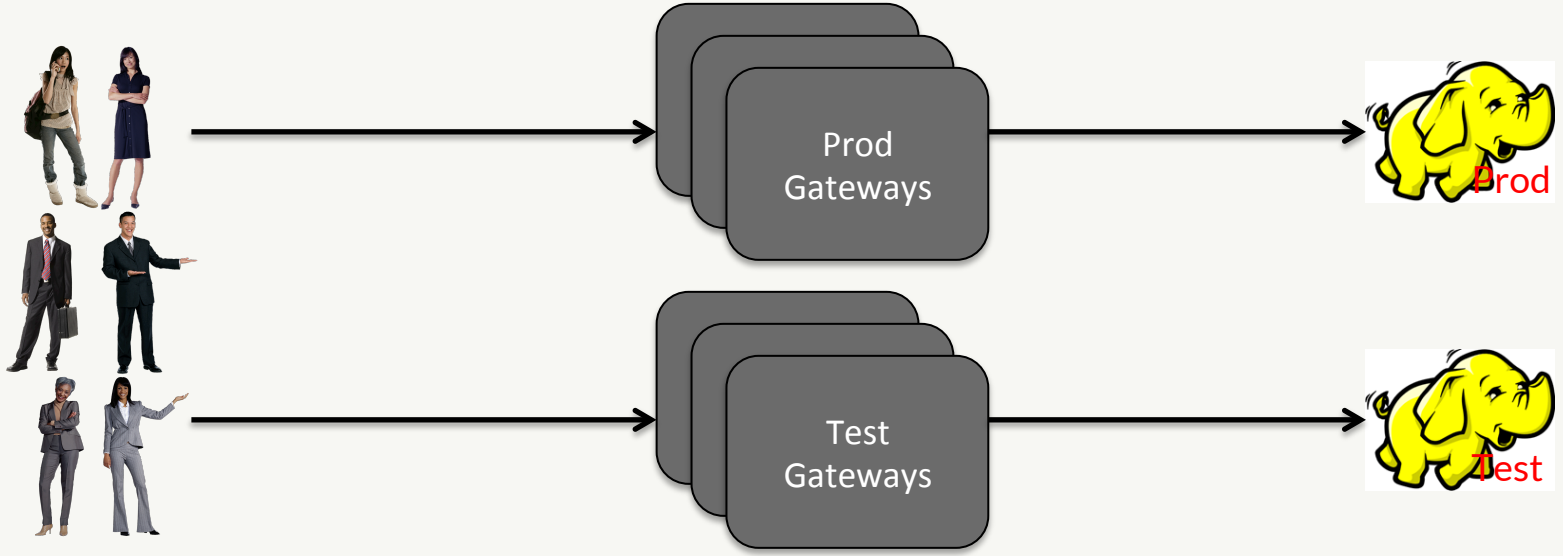
A Nascent Data Platform



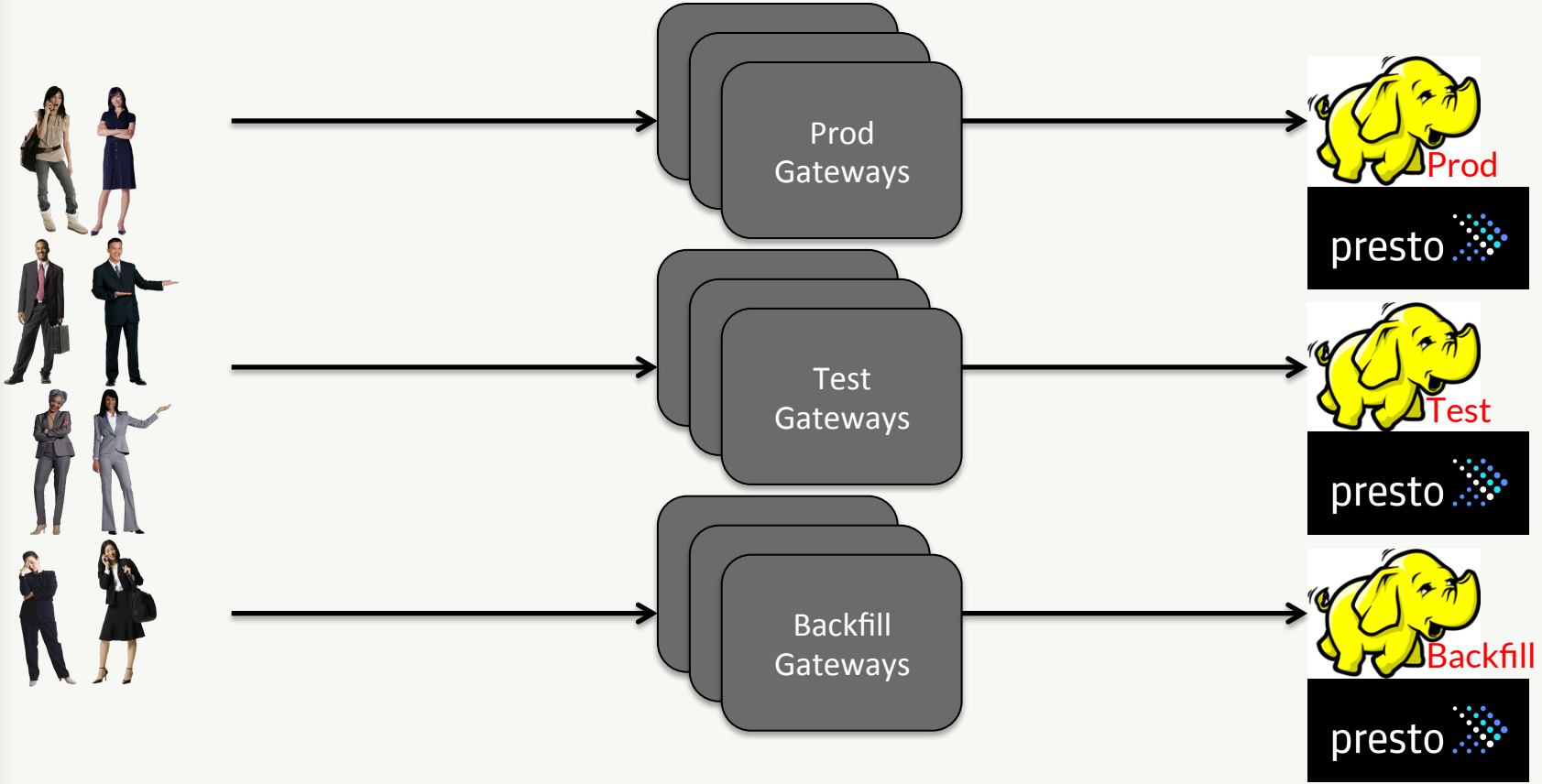
Need Somewhere to Test



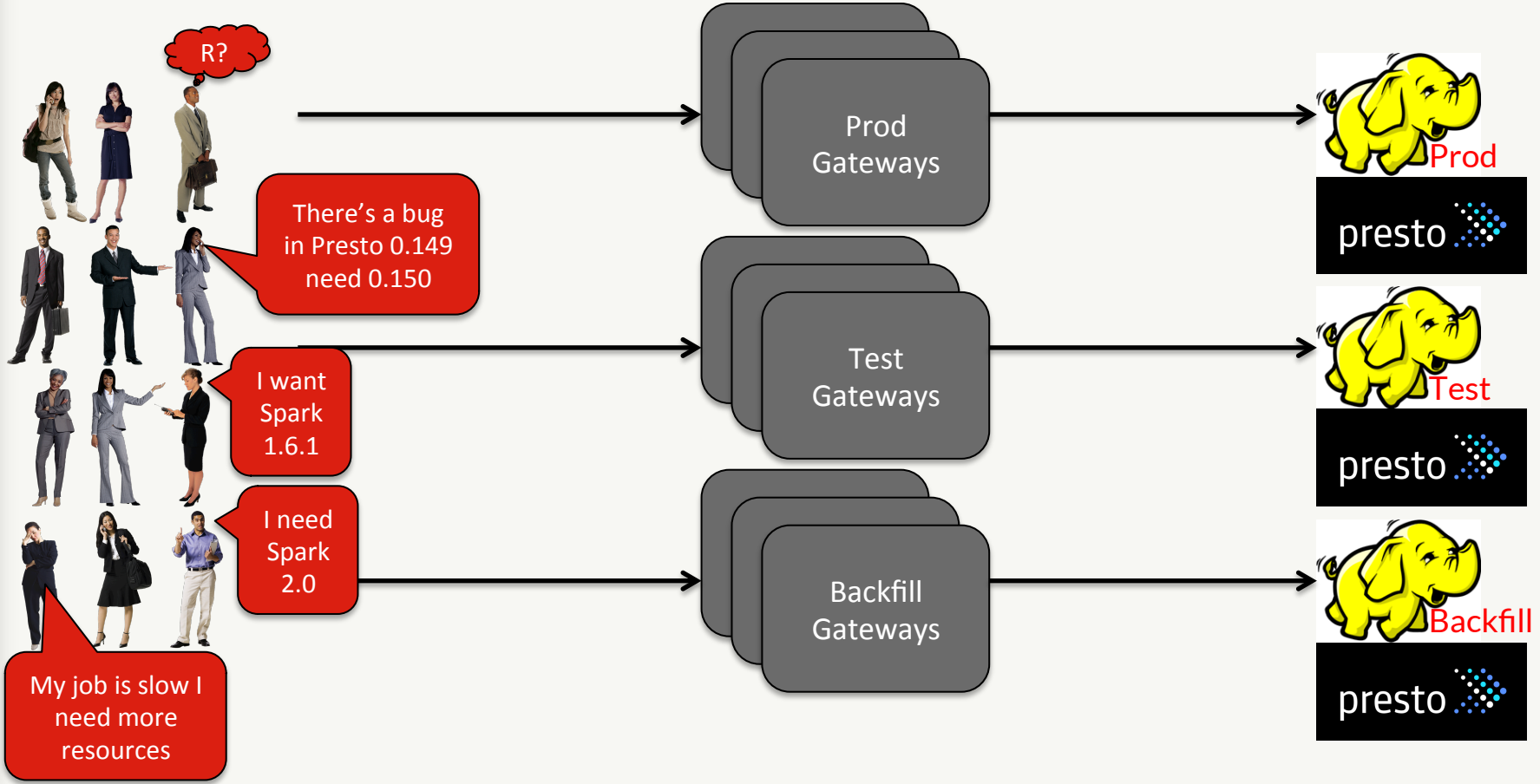
More Users = More Resources

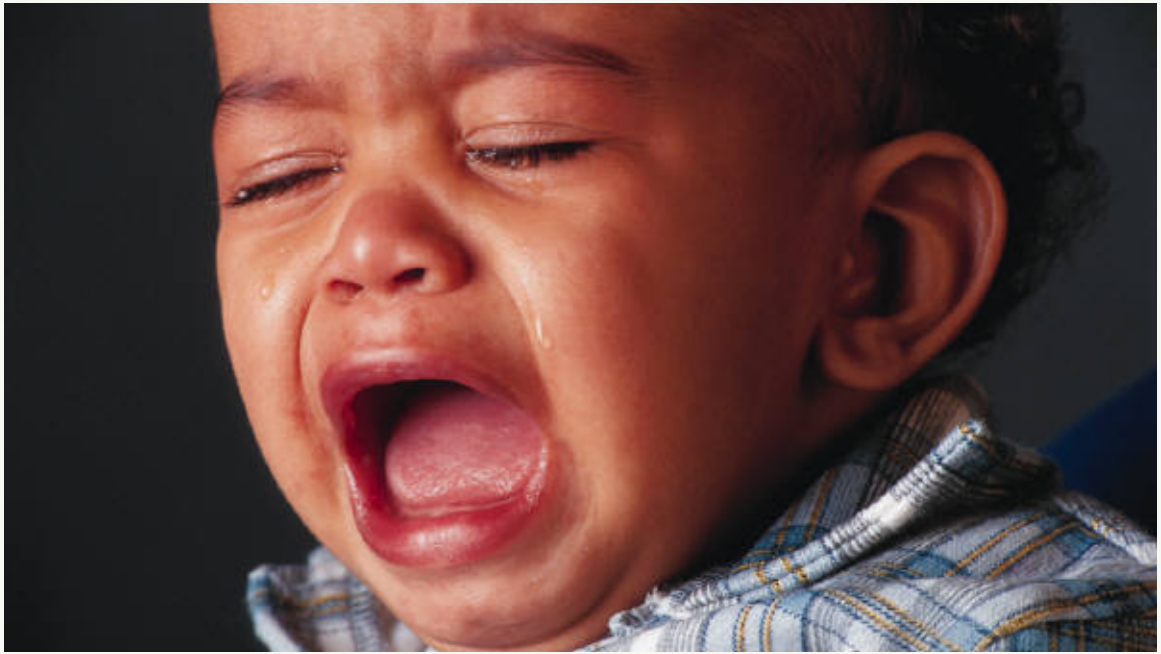


Clusters for Specific Purposes



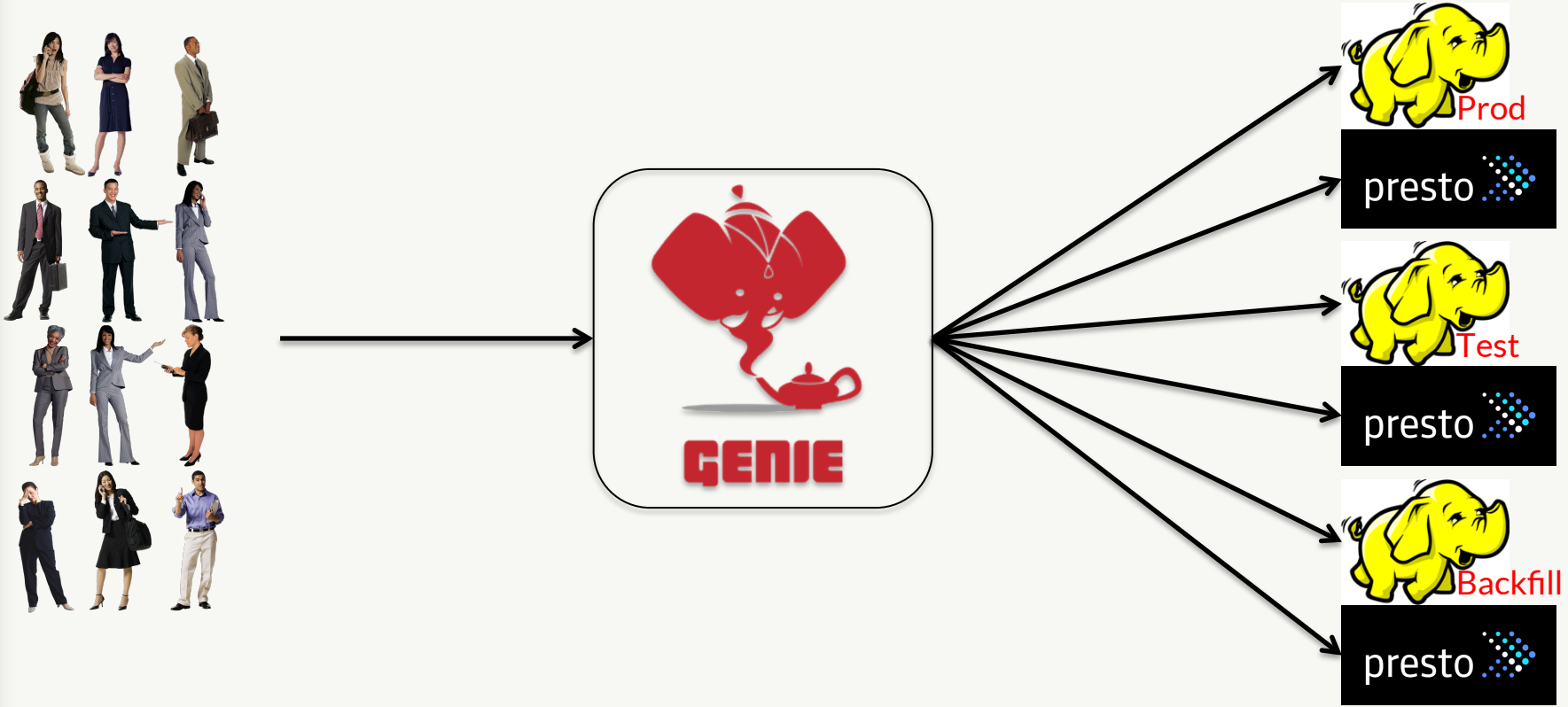
User Base Matures





No one is happy

Genie to the Rescue



Problems Netflix Data Platform Faces

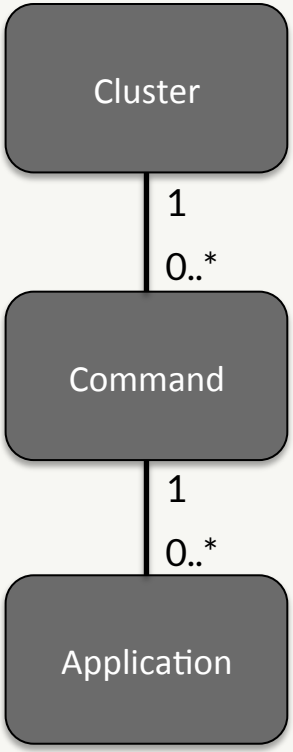
- For Administrators
  - Coordination of many moving parts
    - ~15 clusters
    - ~45 different client executables and versions for those clusters
  - Heavy load
    - ~45-50k jobs per day
  - Hundreds of users with different problems
- For Users
  - Don’t want to know details
  - All clusters and client applications need to be available for use
  - Need to provide tools to make doing their jobs easy

Genie for the Platform Administrator

An administrator wants a tool to…

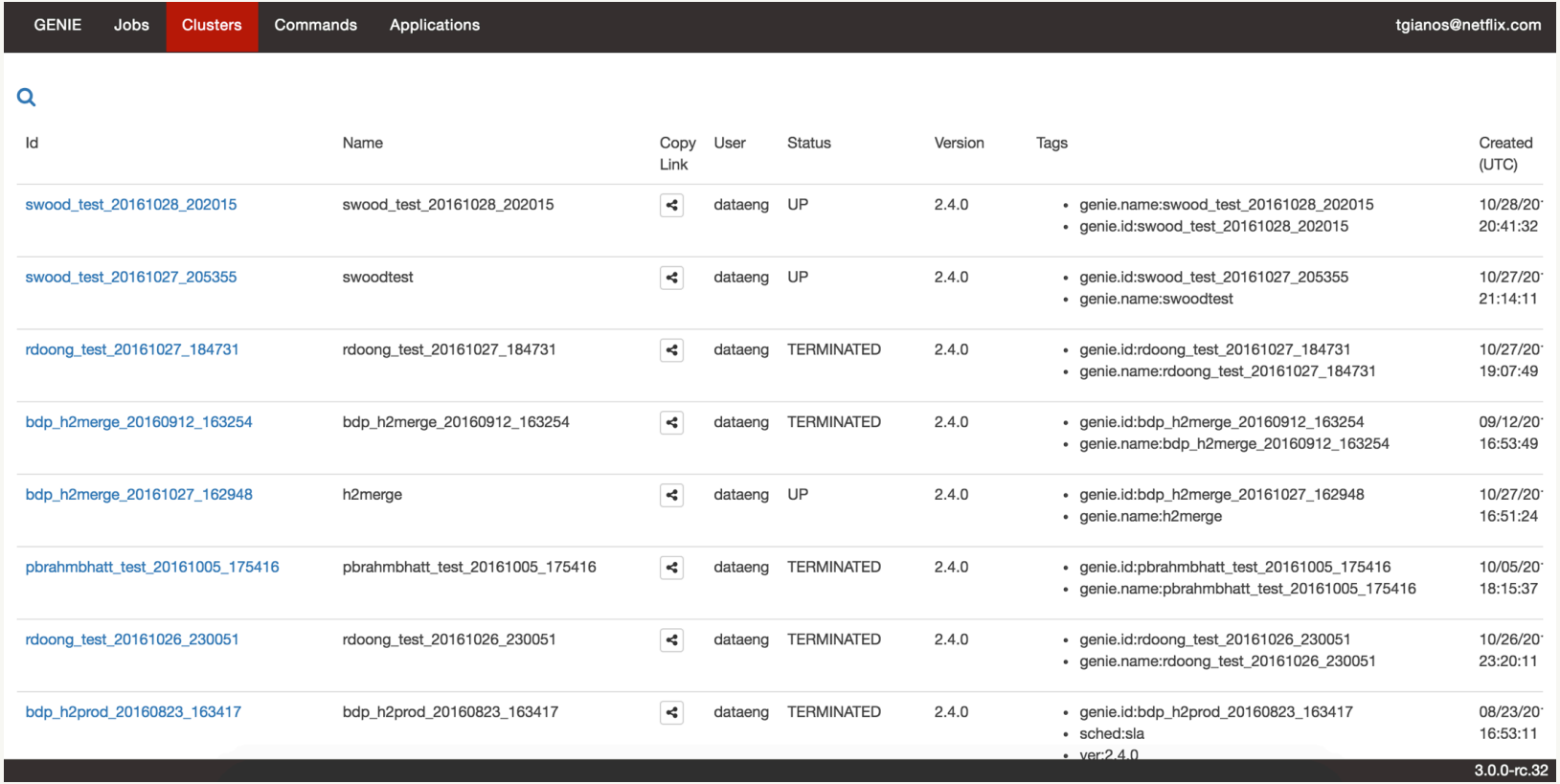
- Simplify configuration management and deployment
- Minimize impact of changes to users
- Track and respond to problems with system quickly
- Scale client resources as load increases

Genie Configuration Data Model



- Metadata about cluster
  - [sched:sla, type:yarn, ver:2.7.1]
- Executable(s)
  - [type:spark-submit, ver:1.6.0]
- Dependencies for an executable

Search Resources



Administration Use Cases

Updating a Cluster

- Start up a new cluster
- Register Cluster with Genie
- Run tests
- Move tags from old to new cluster in Genie
  - New cluster begins taking load immediately
- Let old jobs finish on old cluster
- Shut down old cluster
- No down time!

Load Balance Between Clusters

- Different loads at different times of day
- Copy tags from one cluster to another to split load
- Remove tags when done
- Transparent to all clients!

Update Application Binaries

- Copy new binaries to central download location
- Genie cache will invalidate old binaries on next invocation and download new ones
- Instant change across entire Genie cluster

Genie for Users

User wants a tool to…

- Discover a cluster to run job on
- Run the job client
- Handle all dependencies and configuration
- Monitor the job
- View history of jobs
- Get job results

Submitting a Job

Clusters

```
{
   …
   “clusterCriteria”:[
      “type:yarn”,
      “sched:sla”
   ],
   “commandCriteria”:[
      “type:spark”,
      “ver:1.6.0”
   ]
   …
}
```

GENIE

Spark

Commands

Spark druid presto HIVE sqoop

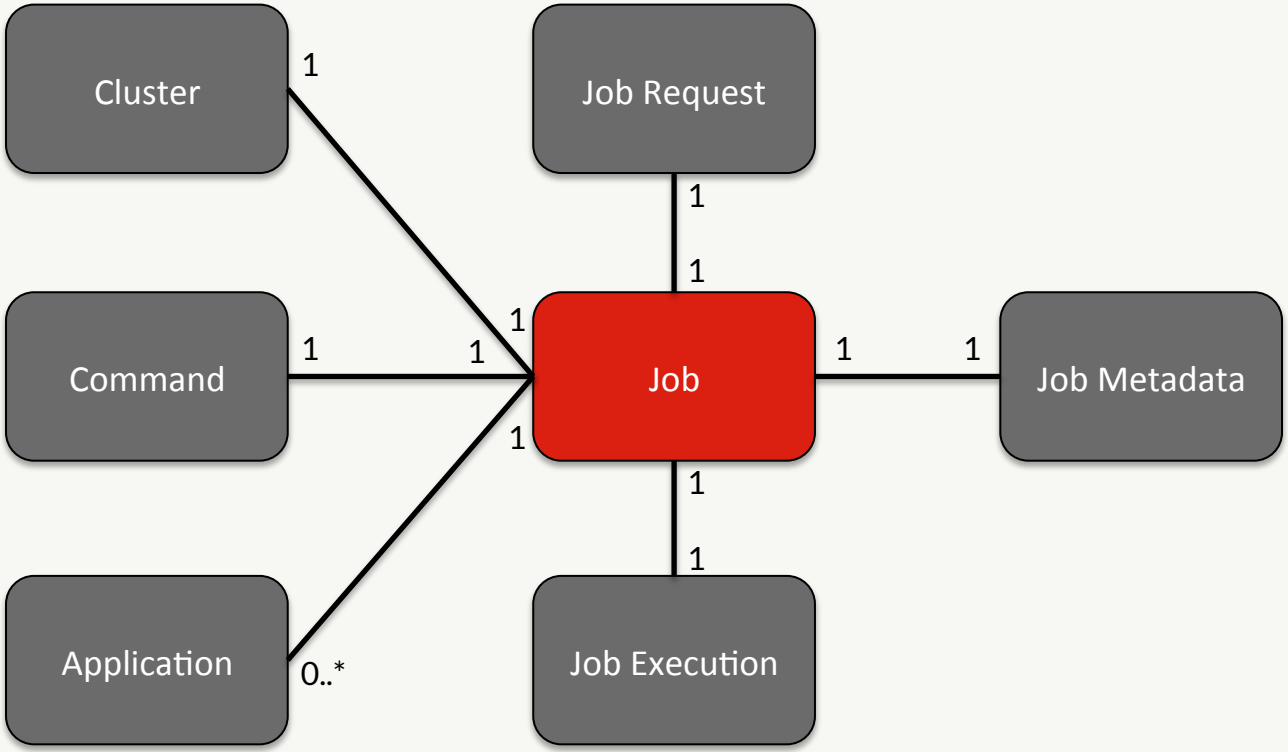
MySQL

presto

presto

1. https://analyticsforinsights.files.wordpress.com/2015/04/superman-data-scientist-graphic.jpg

Genie Job Data Model



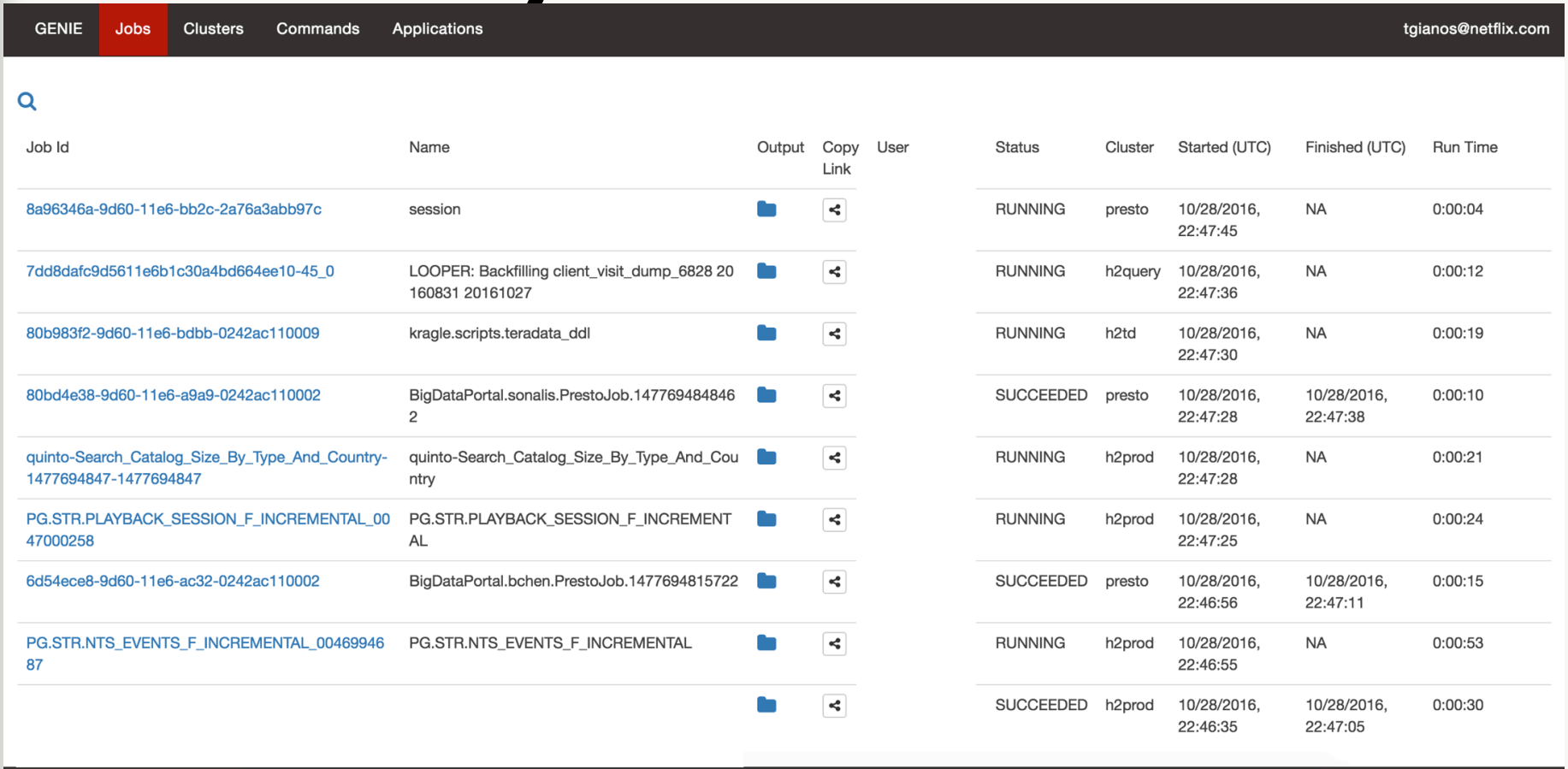
Job Request


```
{
   id: "SPARK.PDA.CLEVENT_F_0046726290",
   created: "2016-10-25T21:26:24.652Z",
   updated: "2016-10-25T21:26:24.652Z",
 - tags: [
       "scheduler.job_name:SPARK.PDA.CLEVENT_F",
       "submitted.by:call_genie",
       "etlcomplete",
       "scheduler.run_id:0046726290",
       "SparkJob",
       "scheduler.name:uc4"
   ],
   version: "NA",
   user: "jasonr",
   name: "SPARK.PDA.CLEVENT_F",
   description: "{"username": "root", "host": "34daee2b00fd", "client": "nflx-kragle-djinn/0.4.1", "kragle_version": "0.40.35", "job_class": "SparkJob"}",
   setupFile: null,
   commandArgs: "--queue root.etlcomplete --driver-memory 8g --num-executors 3000 --class com.netflix.dea.product.cl.clevent_f --executor-memory 6g --conf spark.shuffle.io.maxRetries=10 --conf
   spark.dynamicAllocation.enabled=false --conf spark.io.compression.codec=lz4 --conf spark.hadoop.aws.iam.role.arn=someArnToAssume --conf spark.speculation=false --conf
   spark.yarn.executor.memoryOverhead=3072 --conf spark.driver.maxResultSize=4g --conf spark.sql.shuffle.partitions=9000 DSEPA-product-cl-latest.jar",
 - clusterCriterias: [
     - {
         - tags: [
               "sched:sla"
           ]
       }
   ],
 - commandCriteria: [
       "type:sparksubmit",
       "data:prod"
   ],
   group: null,
   disableLogArchival: false,
   email: null,
   cpu: null,
   memory: null,
   timeout: null,
 - dependencies: [
       "s3://bucket/dea/spark/DSEPA/product-cl/DSEPA-product-cl-latest.jar"
   ],
   applications: [ ],
 - _links: {
     - self: {
           href: "https://genieURL/api/v3/jobs/SPARK.PDA.CLEVENT_F_0046726290/request"
       },
     - job: {
           href: "https://genieURL/api/v3/jobs/SPARK.PDA.CLEVENT_F_0046726290"
       },
     - execution: {
           href: "https://genieURL/api/v3/jobs/SPARK.PDA.CLEVENT_F_0046726290/execution"
       },
     - output: {
           href: "https://genieURL/api/v3/jobs/SPARK.PDA.CLEVENT_F_0046726290/output"
       },
     - status: {
           href: "https://genieURL/api/v3/jobs/SPARK.PDA.CLEVENT_F_0046726290/status"
       }
   }
}
```

Python Client Example

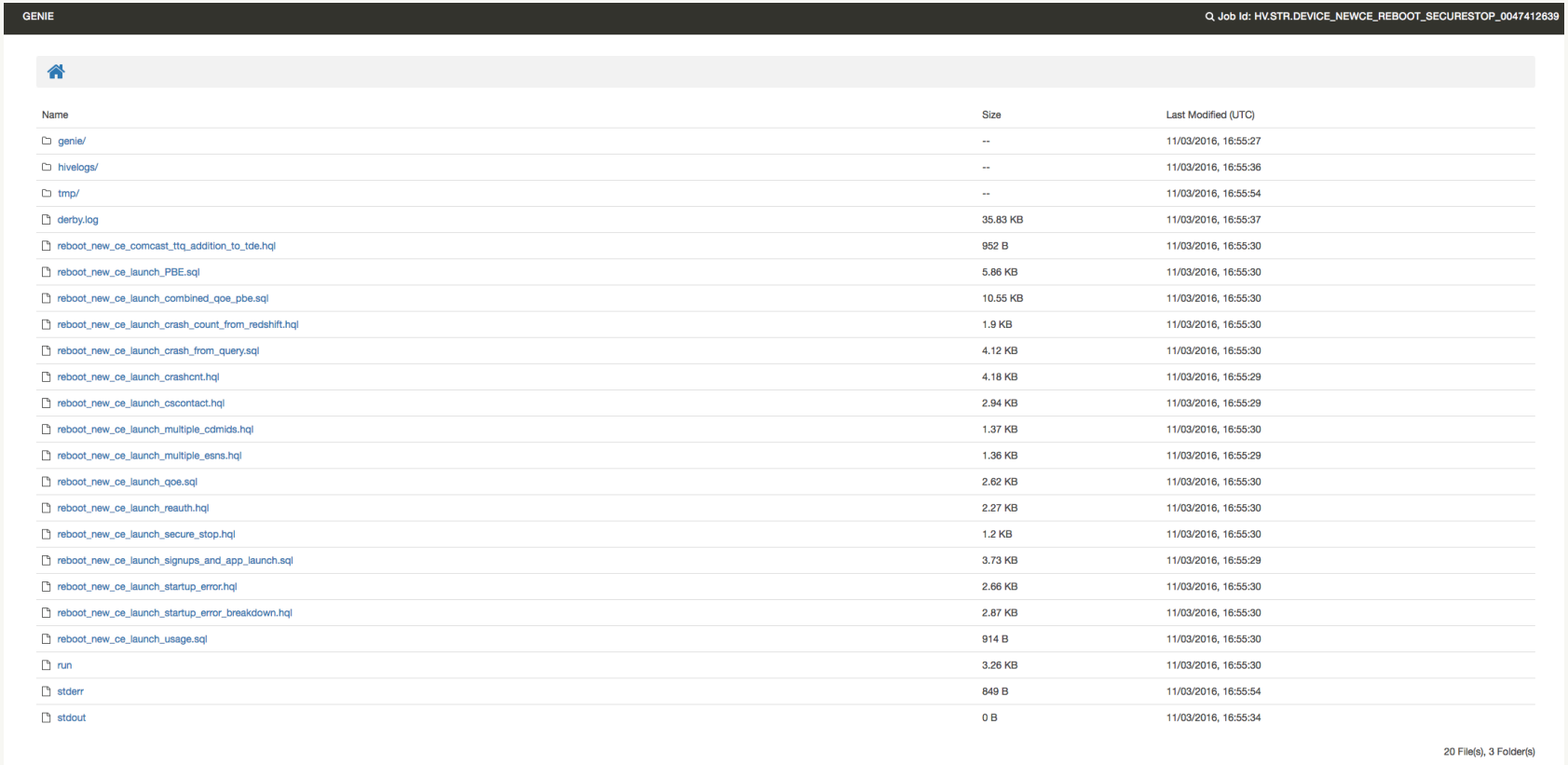
```
import pygenie

job = pygenie.jobs.PrestoJob() \
    .job_name('Presto example') \
    .script("SELECT * FROM my_table WHERE column_1 = '${my_param}'") \
    .parameter('my_param', 'my_param_value') \
    .headers() \                    # include column names in the output
    .option('source', 'examples')    # set --source examples in the command args

# will use default cluster tag "type:presto"
# can override using .cluster_tags() or setting default in config file
# will use default command tag "type:presto"
# can override using .command_tags() or setting default in config file
running_job = job.execute()

print(running_job.job_link)

# block and wait until job is done
running_job.wait()

if not running_job.is_successful:
    print(running_job.stderr())
else:
    print(running_job.stdout())
```

Job History



Job Id	Name	Output	Copy Link	User	Status	Cluster	Started (UTC)	Finished (UTC)	Run Time
8a96346a-9d60-11e6-bb2c-2a76a3abb97c	session				RUNNING	presto	10/28/2016, 22:47:45	NA	0:00:04
7dd8dafc9d5611e6b1c30a4bd664ee10-45_0	LOOPER: Backfilling client_visit_dump_6828 20160831 20161027				RUNNING	h2query	10/28/2016, 22:47:36	NA	0:00:12
80b983f2-9d60-11e6-bdbb-0242ac110009	kragle.scripts.teradata_ddl				RUNNING	h2td	10/28/2016, 22:47:30	NA	0:00:19
80bd4e38-9d60-11e6-a9a9-0242ac110002	BigDataPortal.sonalis.PrestoJob.1477694848462				SUCCEEDED	presto	10/28/2016, 22:47:28	10/28/2016, 22:47:38	0:00:10
quinto-Search_Catalog_Size_By_Type_And_Country-1477694847-1477694847	quinto-Search_Catalog_Size_By_Type_And_Country				RUNNING	h2prod	10/28/2016, 22:47:28	NA	0:00:21
PG.STR.PLAYBACK_SESSION_F_INCREMENTAL_0047000258	PG.STR.PLAYBACK_SESSION_F_INCREMENTAL				RUNNING	h2prod	10/28/2016, 22:47:25	NA	0:00:24
6d54ece8-9d60-11e6-ac32-0242ac110002	BigDataPortal.bchen.PrestoJob.1477694815722				SUCCEEDED	presto	10/28/2016, 22:46:56	10/28/2016, 22:47:11	0:00:15
PG.STR.NTS_EVENTS_F_INCREMENTAL_0046994687	PG.STR.NTS_EVENTS_F_INCREMENTAL				RUNNING	h2prod	10/28/2016, 22:46:55	NA	0:00:53
					SUCCEEDED	h2prod	10/28/2016, 22:46:35	10/28/2016, 22:47:05	0:00:30

Job Output


GENIE
Job Id: HV.STR.DEVICE_NEWCE_REBOOT_SECURESTOP_0047412639
Name
Size
Last Modified (UTC)
genie/ -- 11/03/2016, 16:55:27
hivelogs/ -- 11/03/2016, 16:55:36
tmp/ -- 11/03/2016, 16:55:54
derby.log 35.83 KB 11/03/2016, 16:55:37
reboot_new_ce_comcast_ttq_addition_to_tde.hql 952 B 11/03/2016, 16:55:30
reboot_new_ce_launch_PBE.sql 5.86 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_combined_qoe_pbe.sql 10.55 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_crash_count_from_redshift.hql 1.9 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_crash_from_query.sql 4.12 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_crashcnt.hql 4.18 KB 11/03/2016, 16:55:29
reboot_new_ce_launch_cscontact.hql 2.94 KB 11/03/2016, 16:55:29
reboot_new_ce_launch_multiple_cdmids.hql 1.37 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_multiple_esns.hql 1.36 KB 11/03/2016, 16:55:29
reboot_new_ce_launch_qoe.sql 2.62 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_reauth.hql 2.27 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_secure_stop.hql 1.2 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_signups_and_app_launch.sql 3.73 KB 11/03/2016, 16:55:29
reboot_new_ce_launch_startup_error.hql 2.66 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_startup_error_breakdown.hql 2.87 KB 11/03/2016, 16:55:30
reboot_new_ce_launch_usage.sql 914 B 11/03/2016, 16:55:30
run 3.26 KB 11/03/2016, 16:55:30
stderr 849 B 11/03/2016, 16:55:54
stdout 0 B 11/03/2016, 16:55:34
20 File(s), 3 Folder(s)

Wrapping Up

Data Warehouse

- S3 for Scale
- Decouple Compute & Storage
- Parquet for Speed

Genie at Netflix

- Runs the OSS code
- Runs ~45k jobs per day in production
- Runs on ~25 i2.4xl instances at any given time
- Keeps ~3 months of jobs (~3.1 million) in history

Resources

- http://netflix.github.io/genie/
  - Work in progress for 3.0.0
- https://github.com/Netflix/genie
  - Demo instructions in README
- https://hub.docker.com/r/netflixoss/genie-app/
  - Docker Container

Questions?