Synthetic Data Generation for Realistic Analytics Examples and Testing

Ronald J. Nowling
Red Hat, Inc.
rnowling@redhat.com
http://rnowling.github.io/

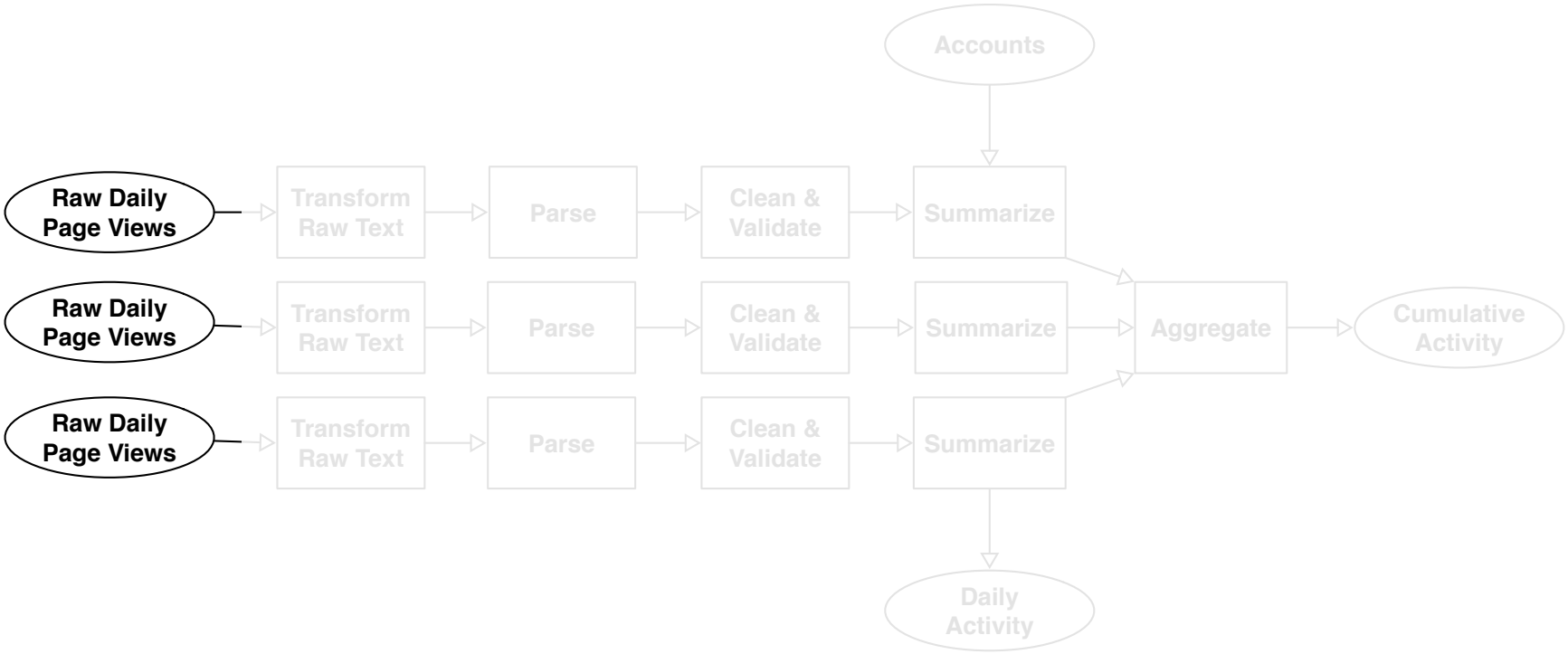
Who Am I?

- Software Engineer at Red Hat
- Data Science Team, Emerging Technologies
  - Evaluate open-source Big Data space
  - Ensure software works for Red Hat customers
  - Promote data science internally through consulting projects
- Apache BigTop PMC

Synthetic Data

- No licensing, privacy, or intellectual property concerns
- Scalable: Laptops to Clusters!
- More reliable than external data sets
- Enable more realistic example applications
- Enable more comprehensive testing than wordcount and TeraSort

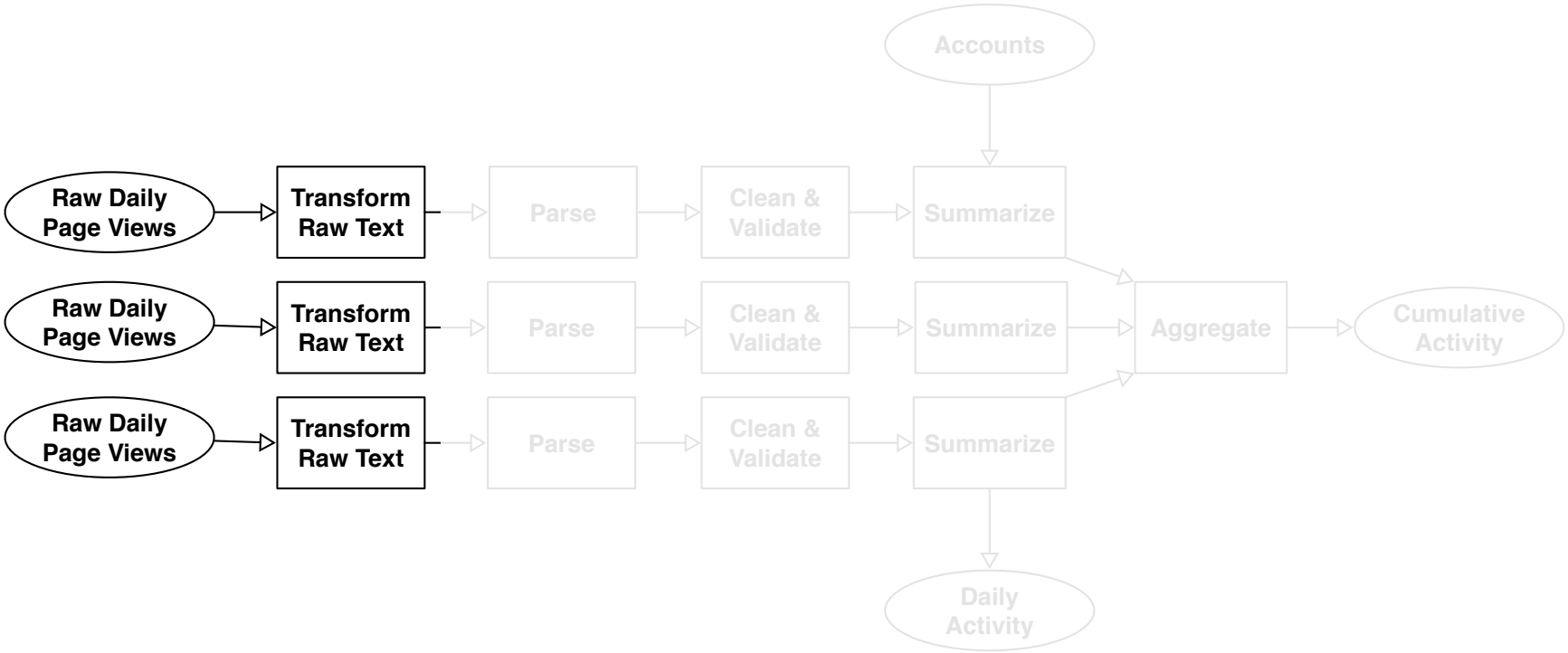
Data Transformation and Summarization Pipeline



redhat

Spark

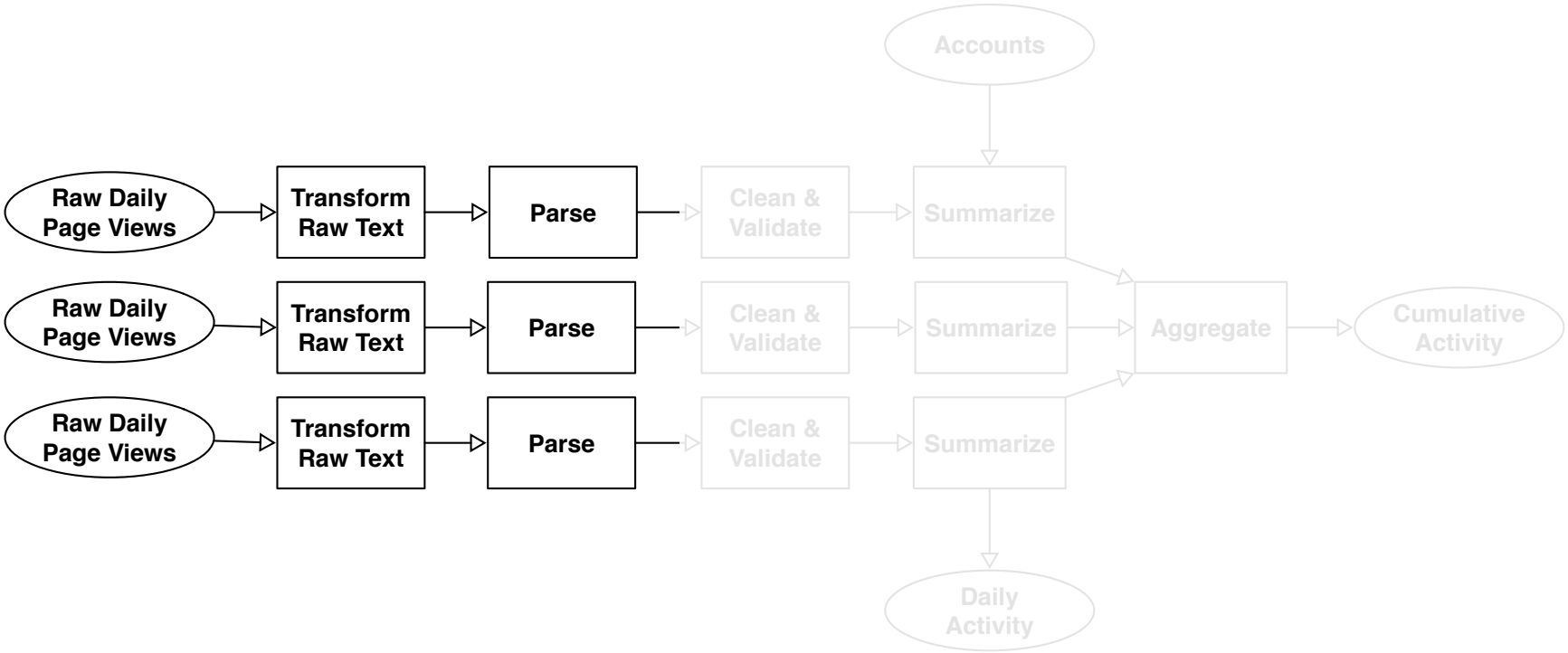
Data Transformation and Summarization Pipeline



redhat

Spark

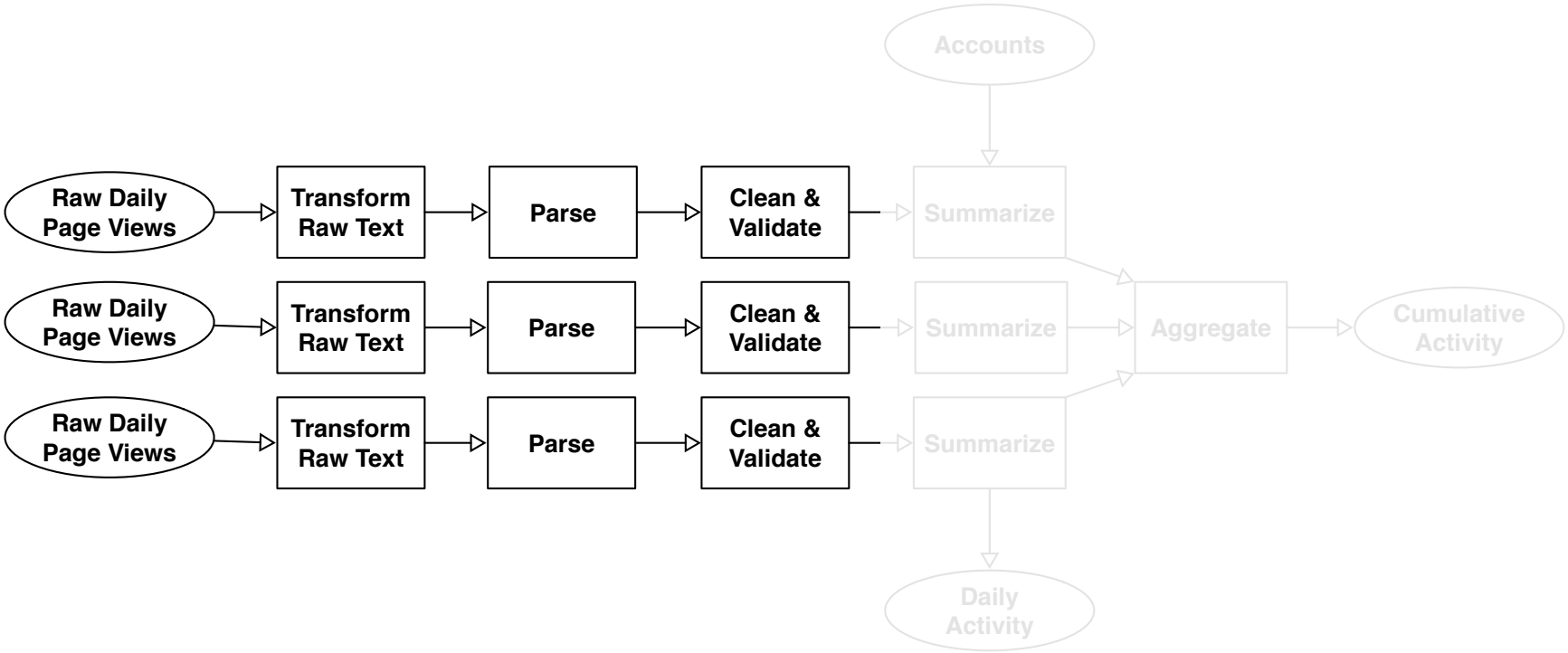
Data Transformation and Summarization Pipeline



redhat

Spark

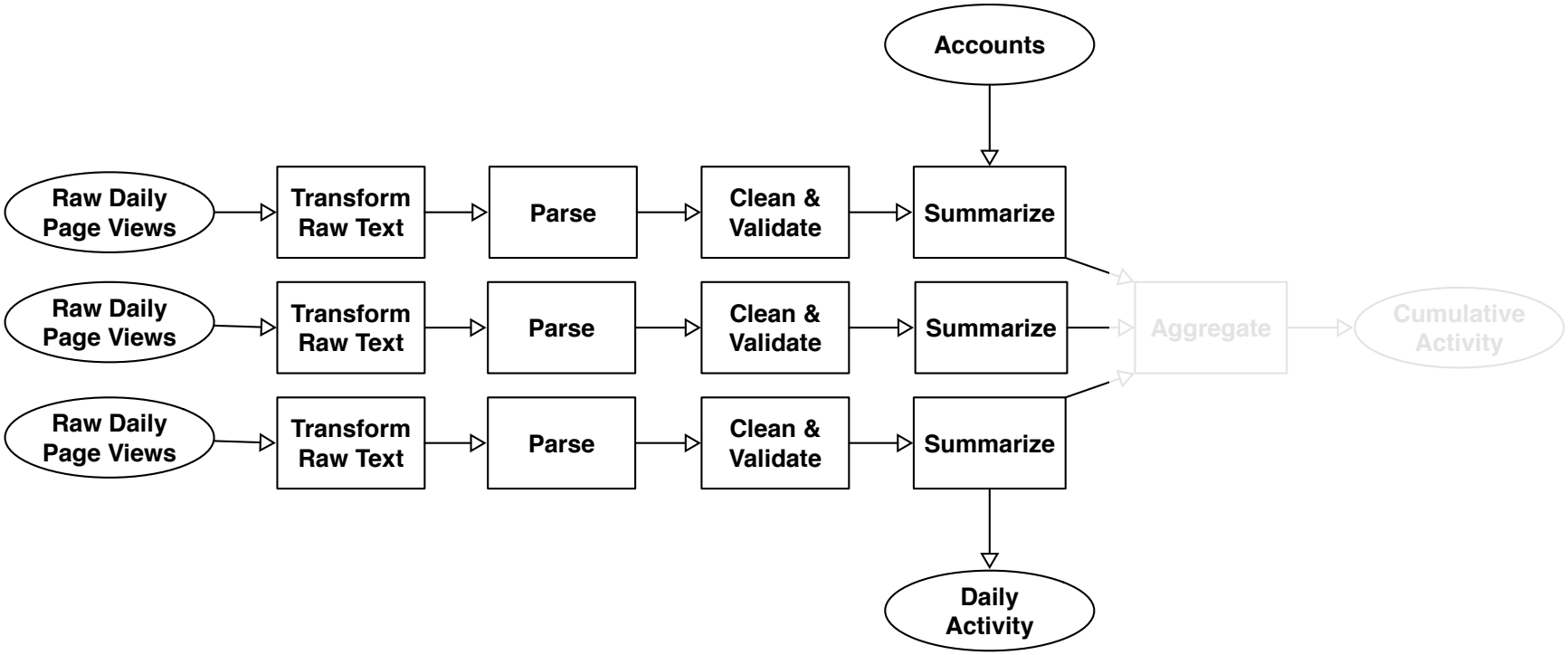
Data Transformation and Summarization Pipeline



redhat

Spark

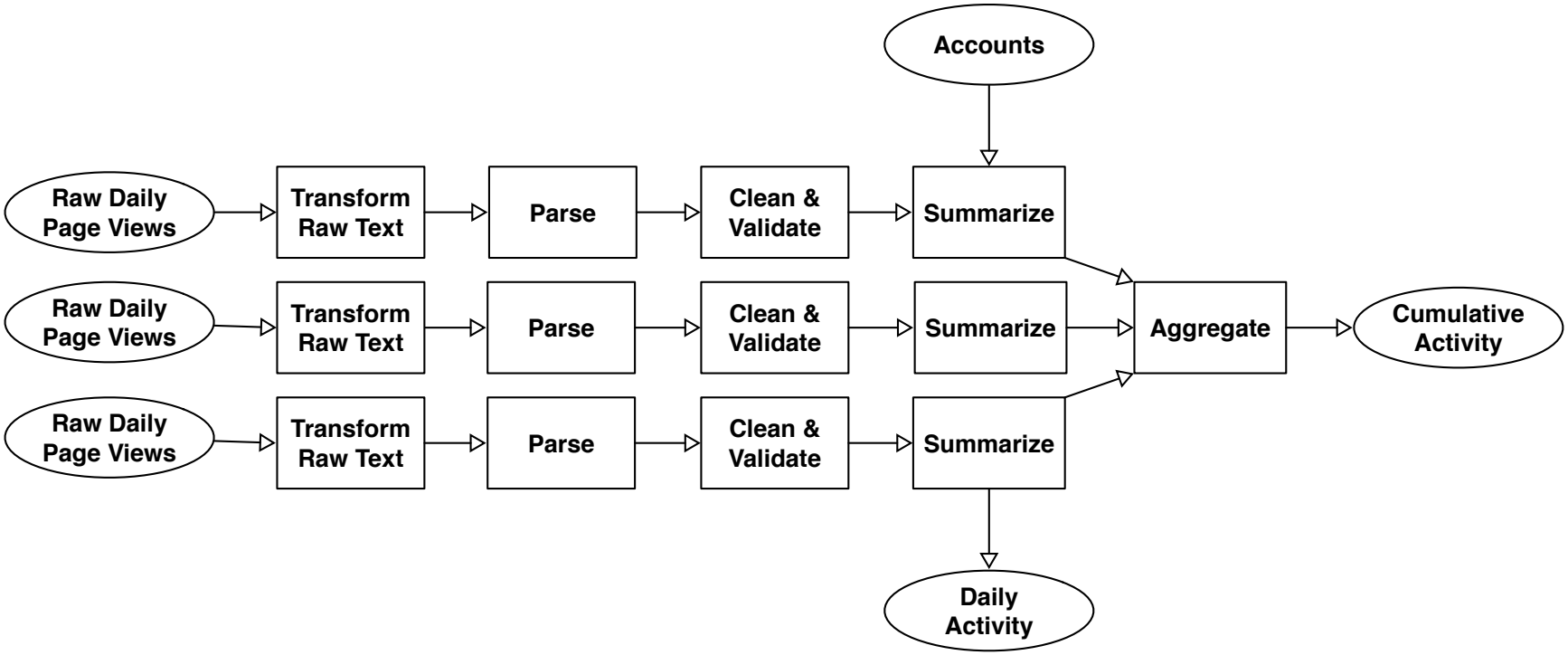
Data Transformation and Summarization Pipeline



redhat

Spark

Data Transformation and Summarization Pipeline



redhat

Spark

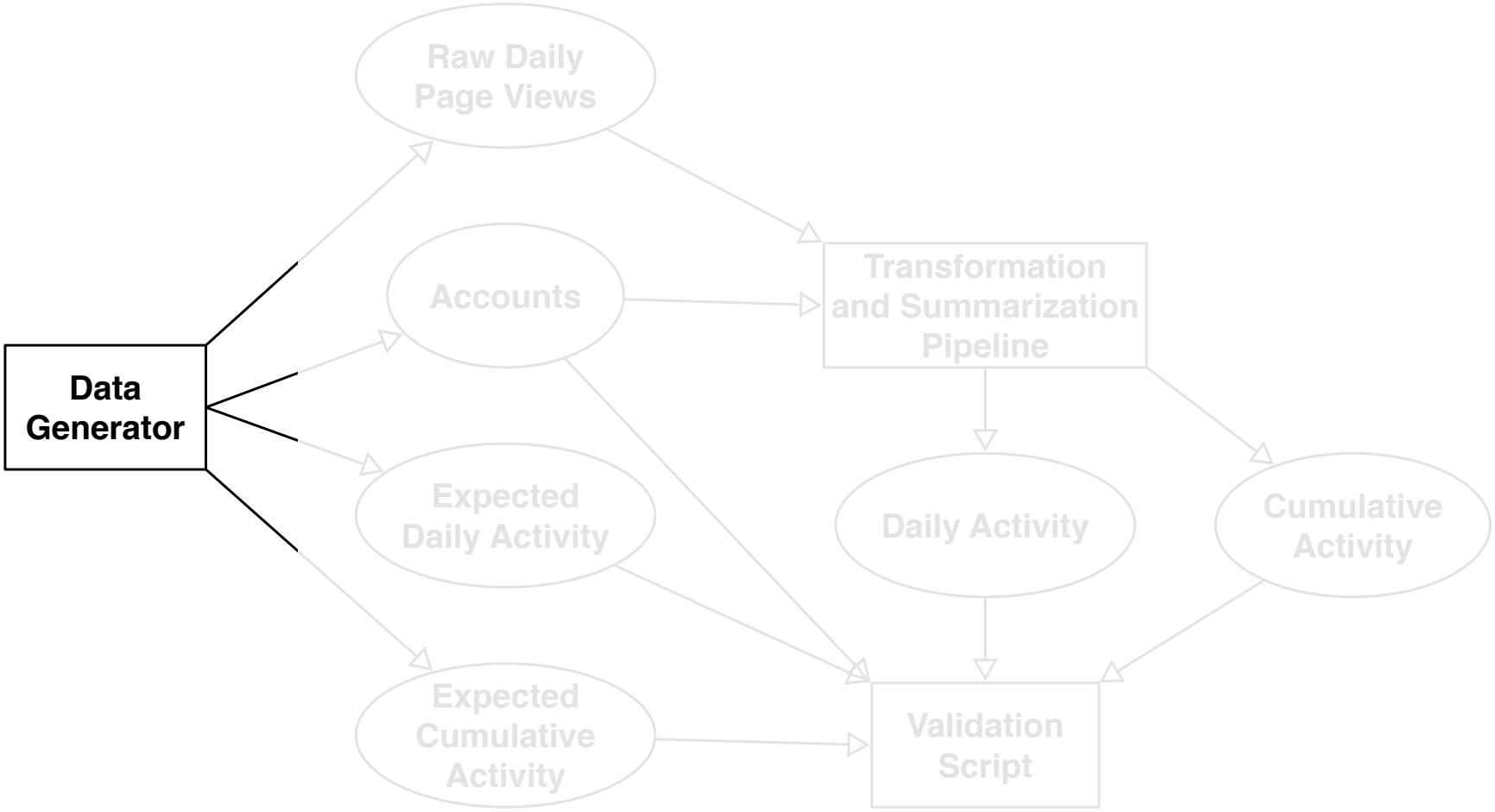
Timings

- Data set
  - 1000's of files
  - 100's of GBs compressed (gzip)
- Conversion from .tsv.gz -> Parquet
  ~45 min
- Compute aggregations on Parquet data and write out
  ~2 min

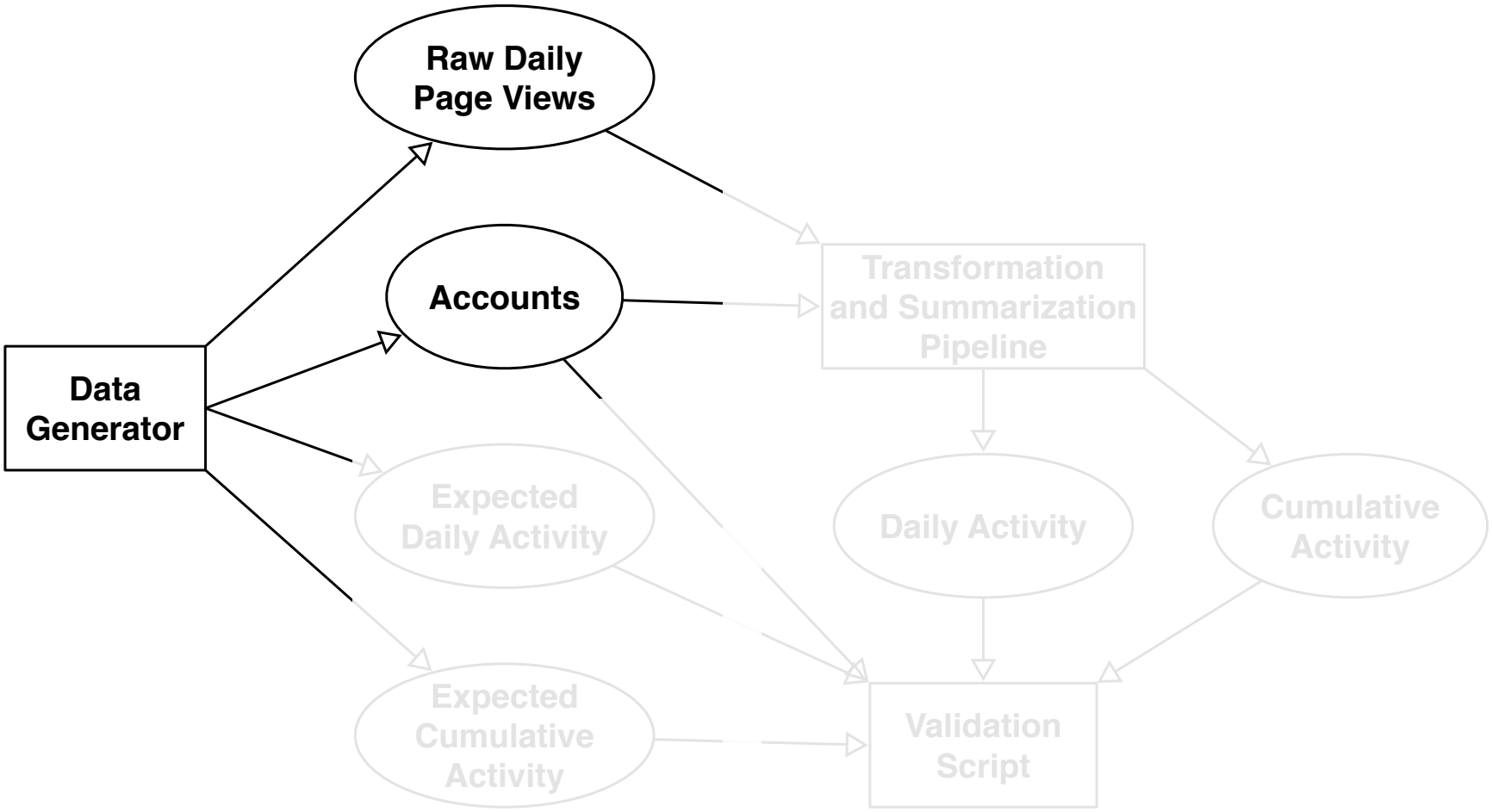
Synthetic Data

- Sensitive Data
  - Real data on cluster for scalability testing and validation
  - Synthetic data for local development and testing
- Smaller data sets for checking calculations
  - Total aggregation results requires re-running old pipeline
  - Extra burden on operations team
  - Delay for development team

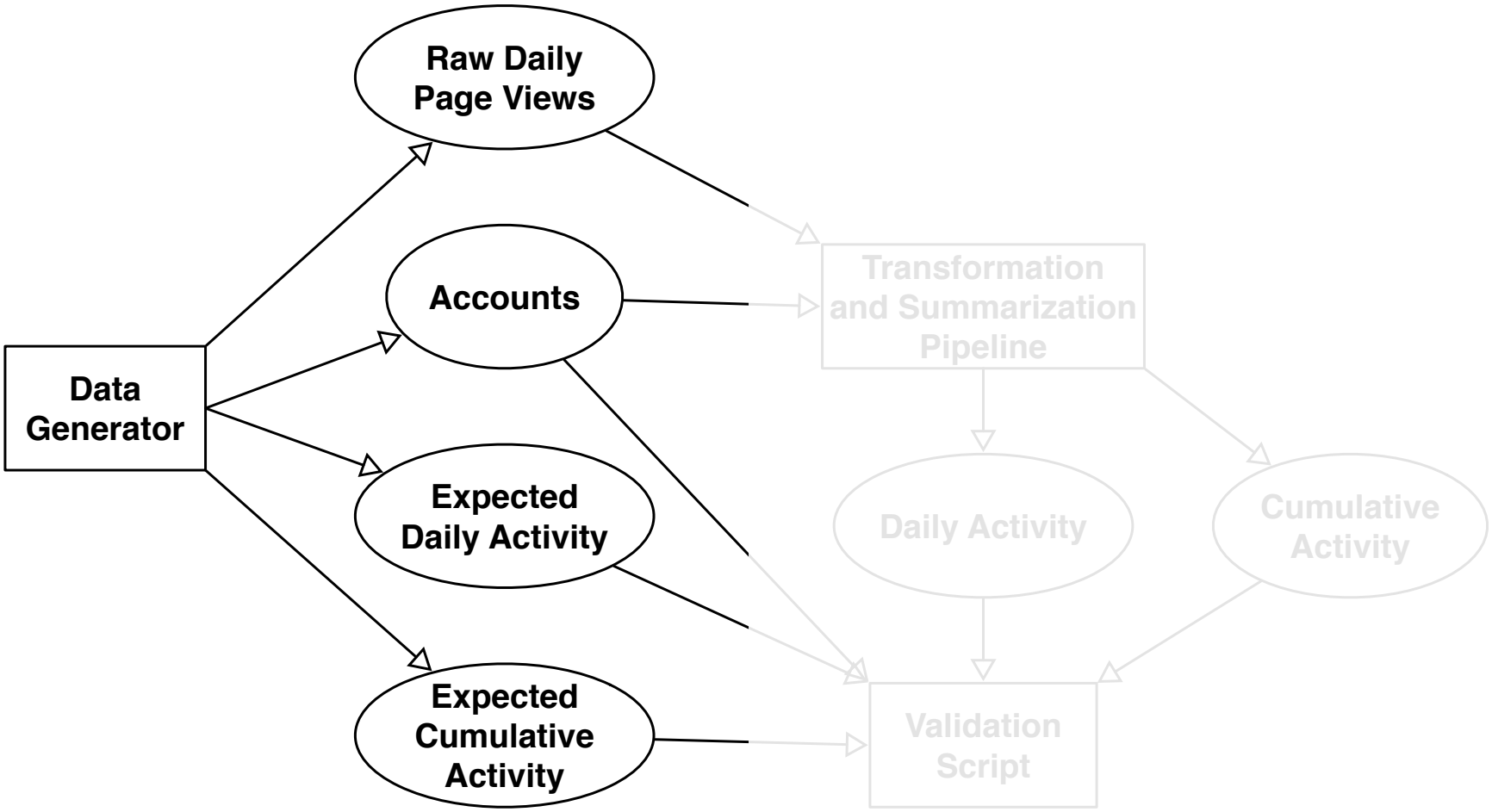
Validation with Synthetic Data



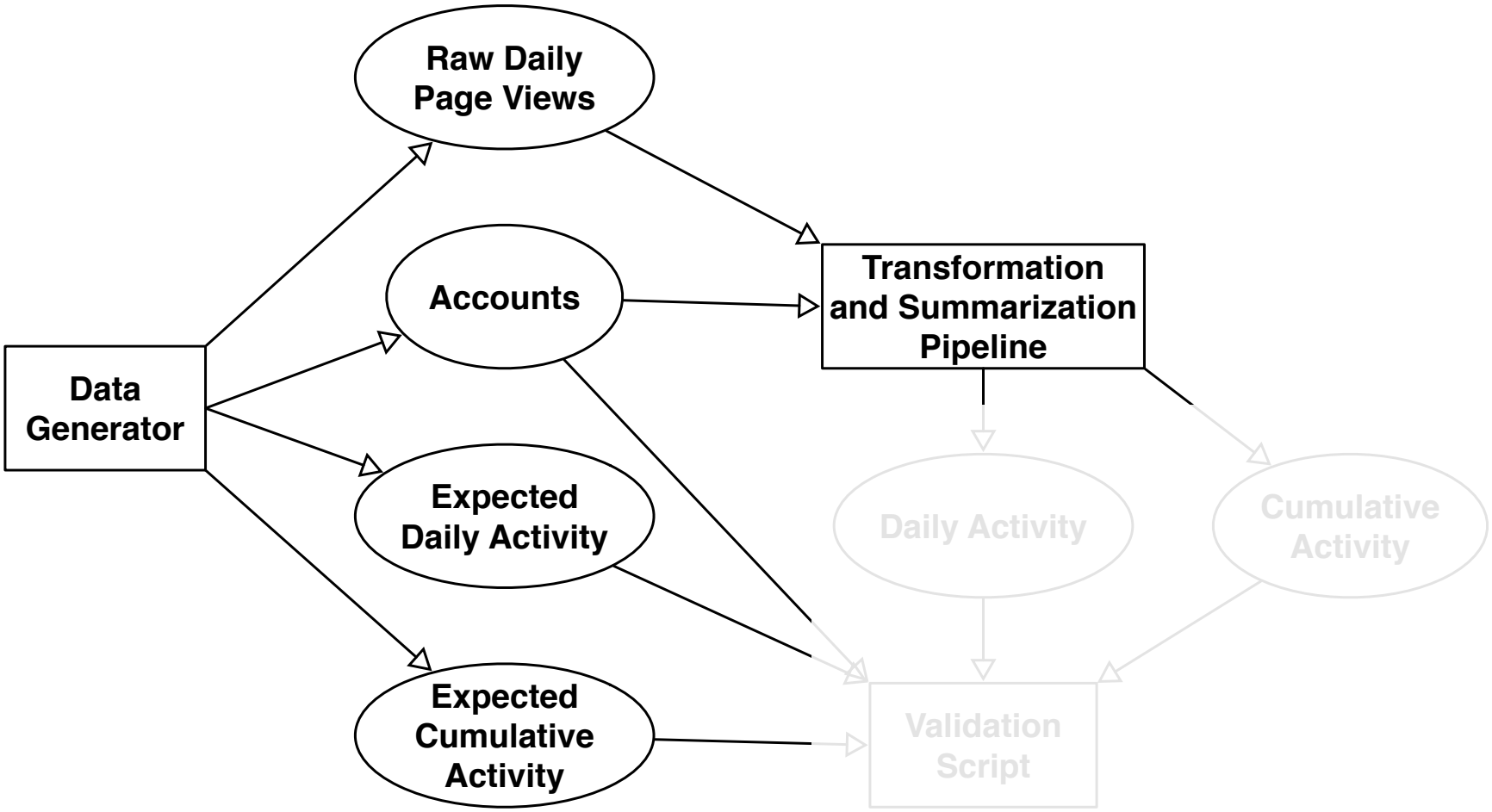
Validation with Synthetic Data



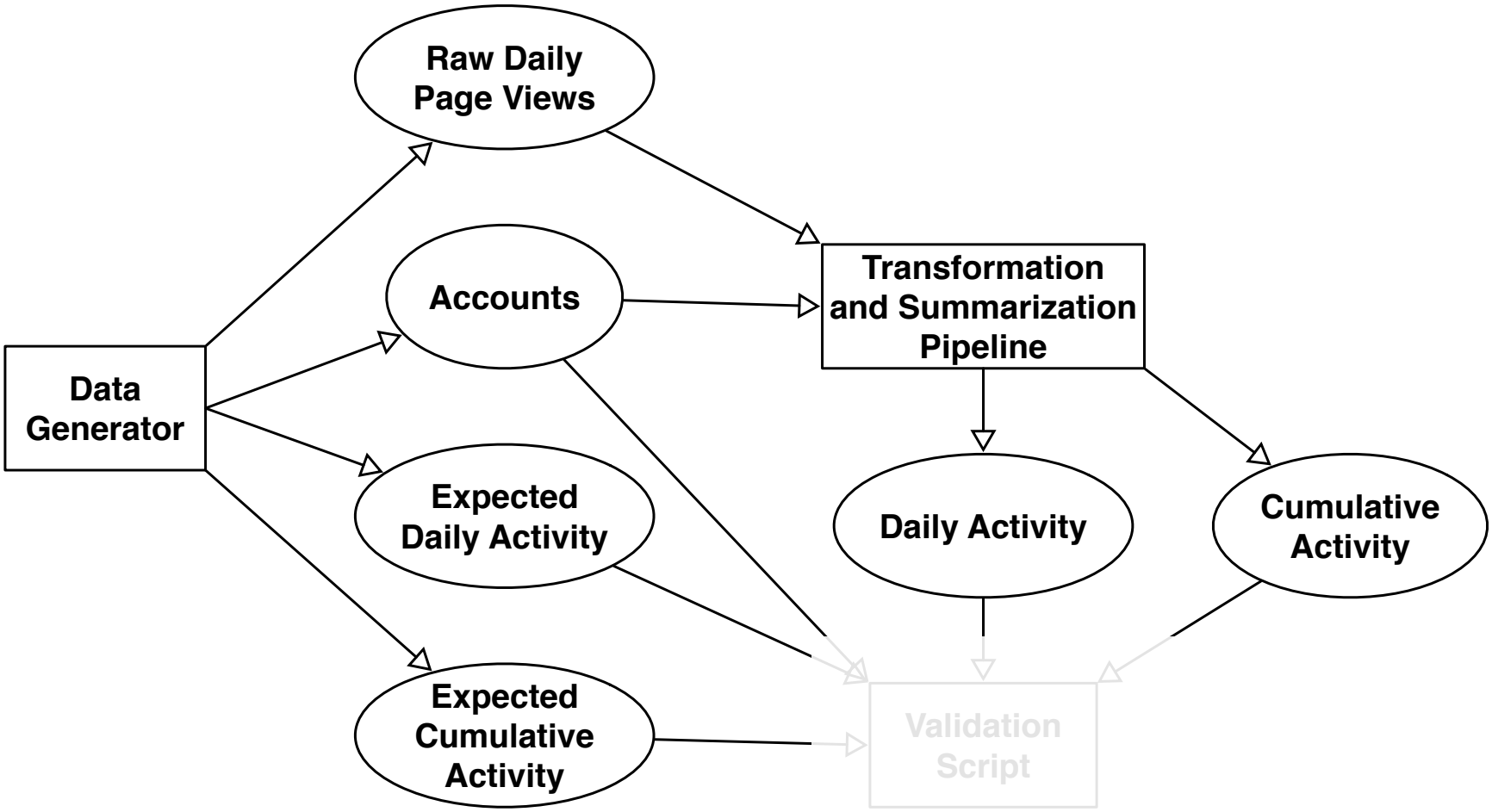
Validation with Synthetic Data



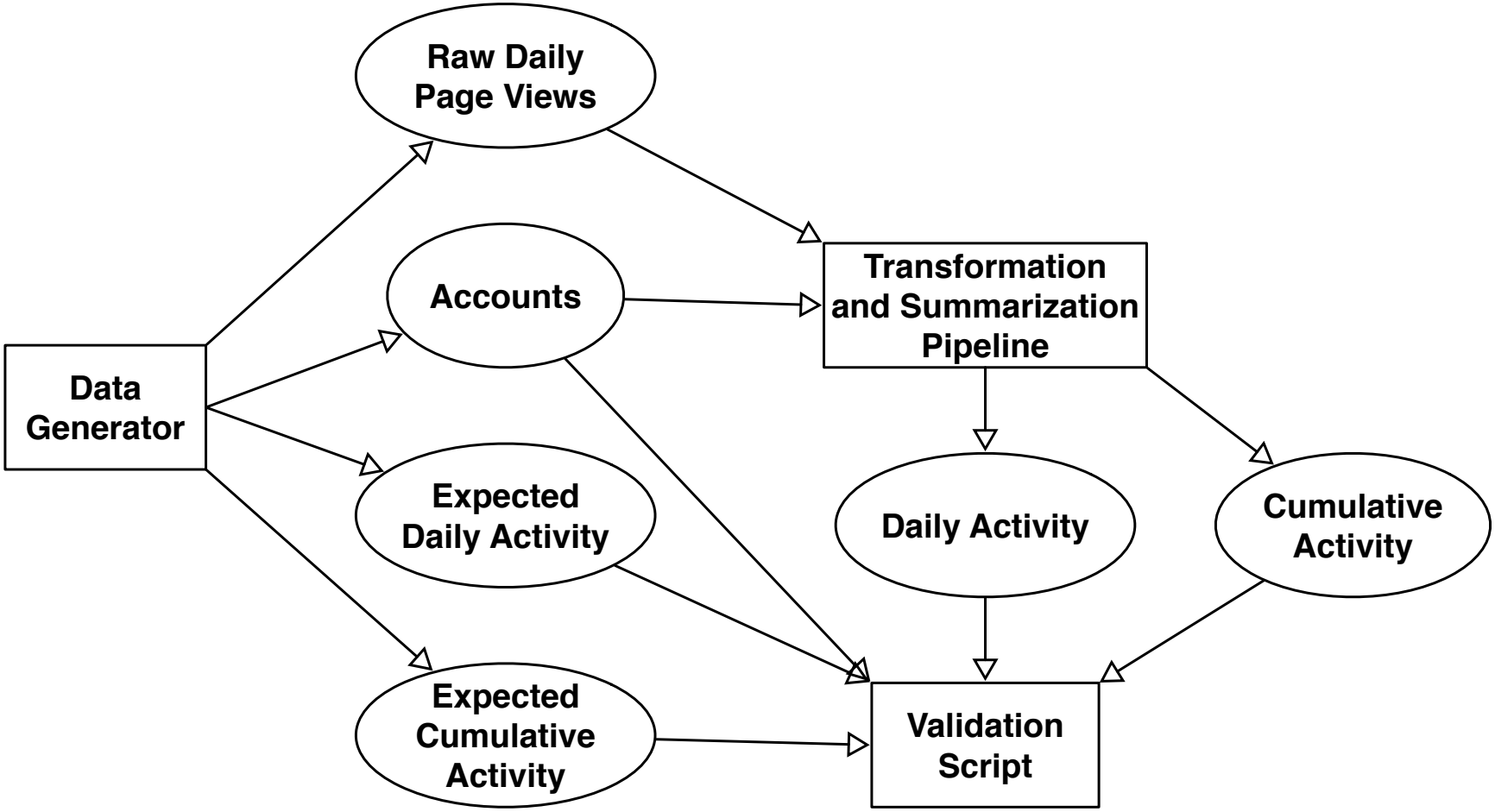
Validation with Synthetic Data



Validation with Synthetic Data



Validation with Synthetic Data



Issues Tackled

- Error in account validation introduced while refactoring code
- Usage of the correct join types
- Validation of date-time operations
- Correct Output Formats

Gzipped Files

- Gzip doesn't support random access – entire file needs to be decompressed sequentially
- Large files – multiple gigabytes uncompressed
- Too many files read in parallel –> long GC or OOM errors

(Quirky) TSV Files

- Tab-separated, no quoting
- Escaped tabs and newlines within records
  - E.g., \\n or \\t
- Improperly escaped tabs and newlines
  - E.g., \\\t vs \\\\t

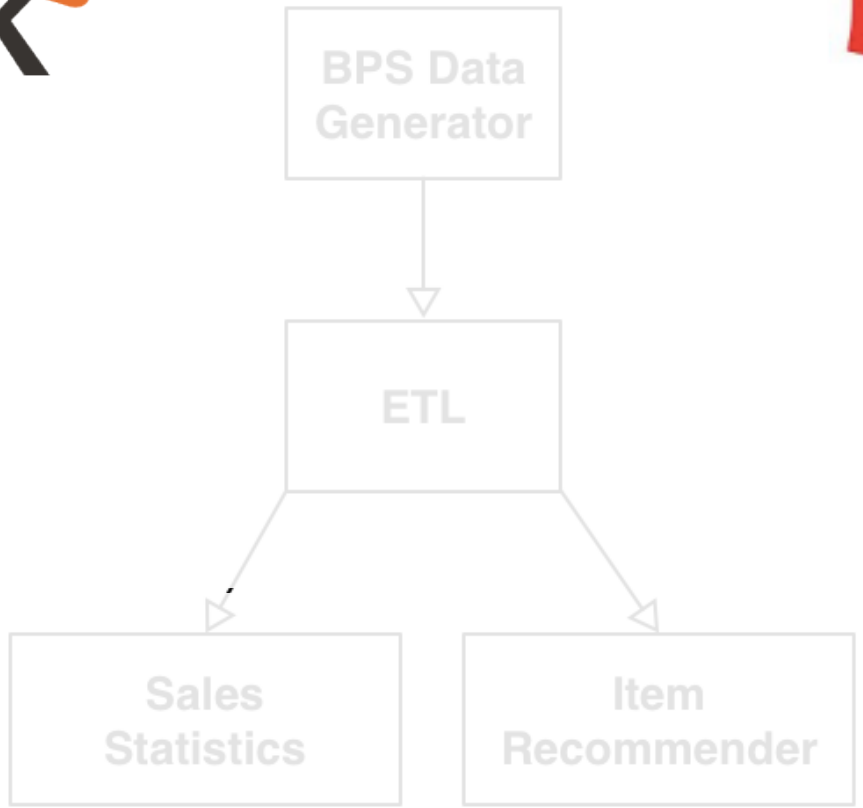
Solutions

- Convert to Parquet as quickly as possible
- Use fewer cores per node
  - More RAM / task (partition)
- 2-phase grouping algorithm
  - Group within partition
  - Group partition ends using shuffle
  - Union
- Optimized string operations
  - Use iterators instead of concatenation and replace
  - Custom CSV parser implementation

Apache BigTop BigPetStore Blueprints

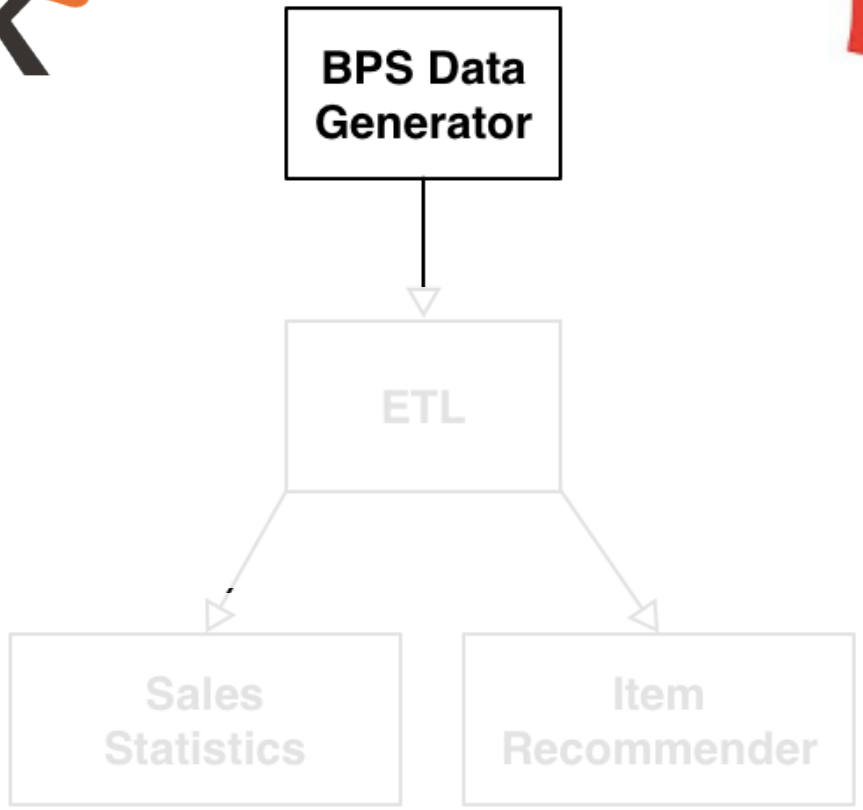
- Problem domain: Transactions for a fictional chain of pet stores
- BigPetStore Data Generator simulates customer purchasing behavior to generate realistic transaction data
- Blueprints for big data ecosystem
  - Hadoop: MapReduce / Pig / Hive / Mahout
  - Spark
  - Flink (in progress)

BigPetStore



BigPetStore

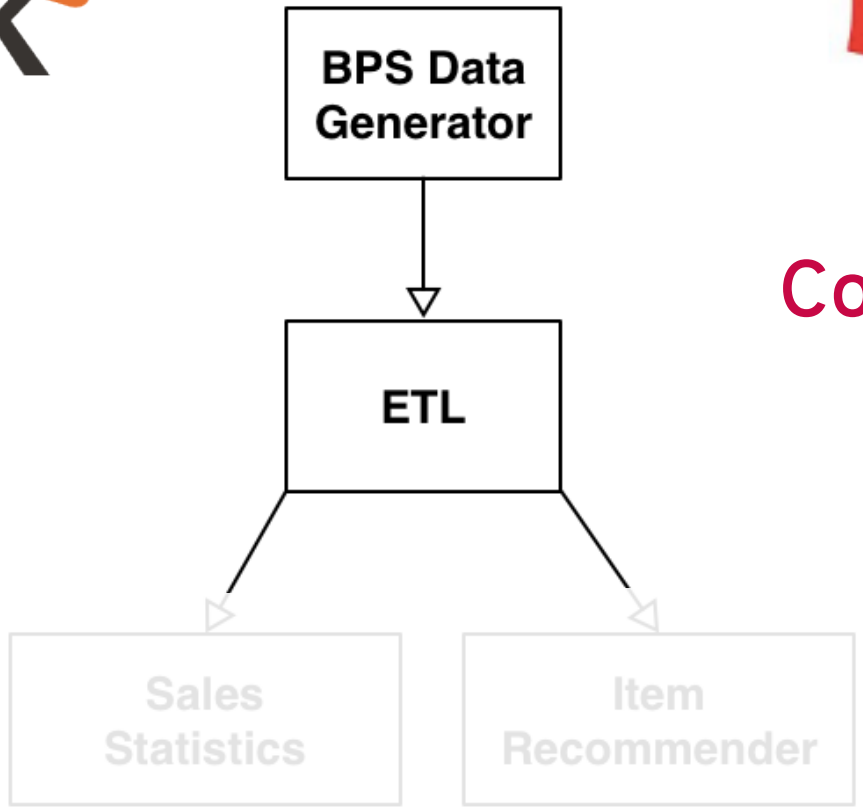
HCFS



BigPetStore

HCFS

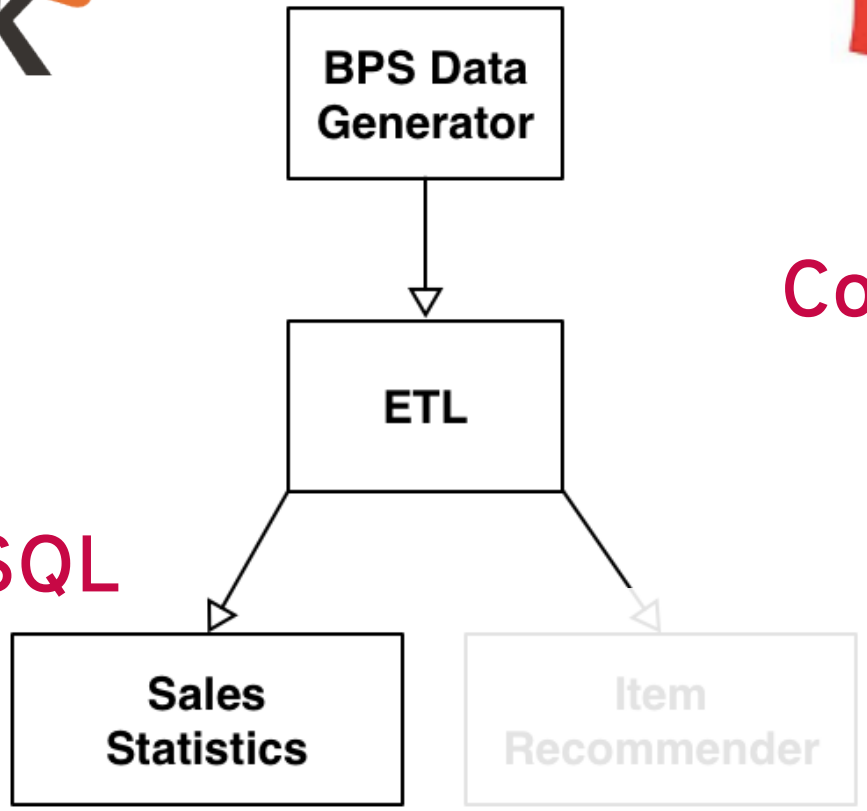
Core (RDDs)



BigPetStore

HCFS

Core (RDDs)

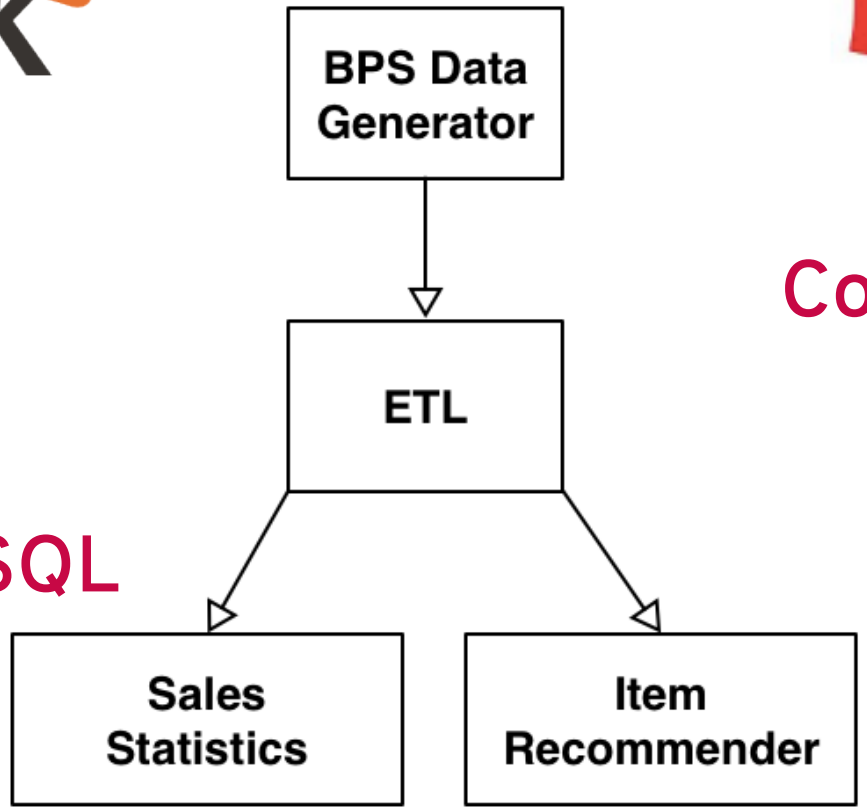


Spark SQL

BigPetStore

HCFS

Core (RDDs)



Spark SQL

MLLib

Team Cluster

- ~10 nodes
- 40 cores, 400GB RAM per node

ANSIBLE

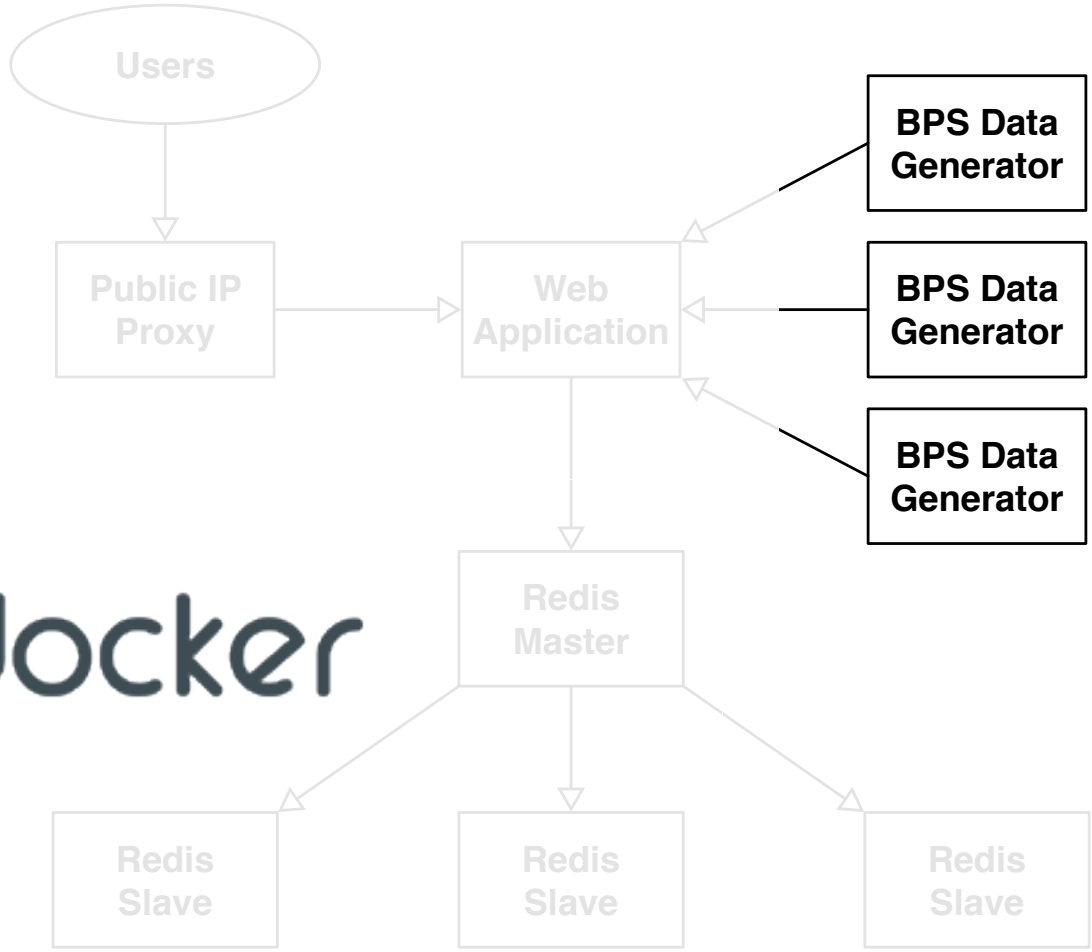
Potential Issues

- Infrastructure
- Storage
- Software Installation
- Software Upgrades
- Spark Configuration Tuning
- User Management

Real Stories

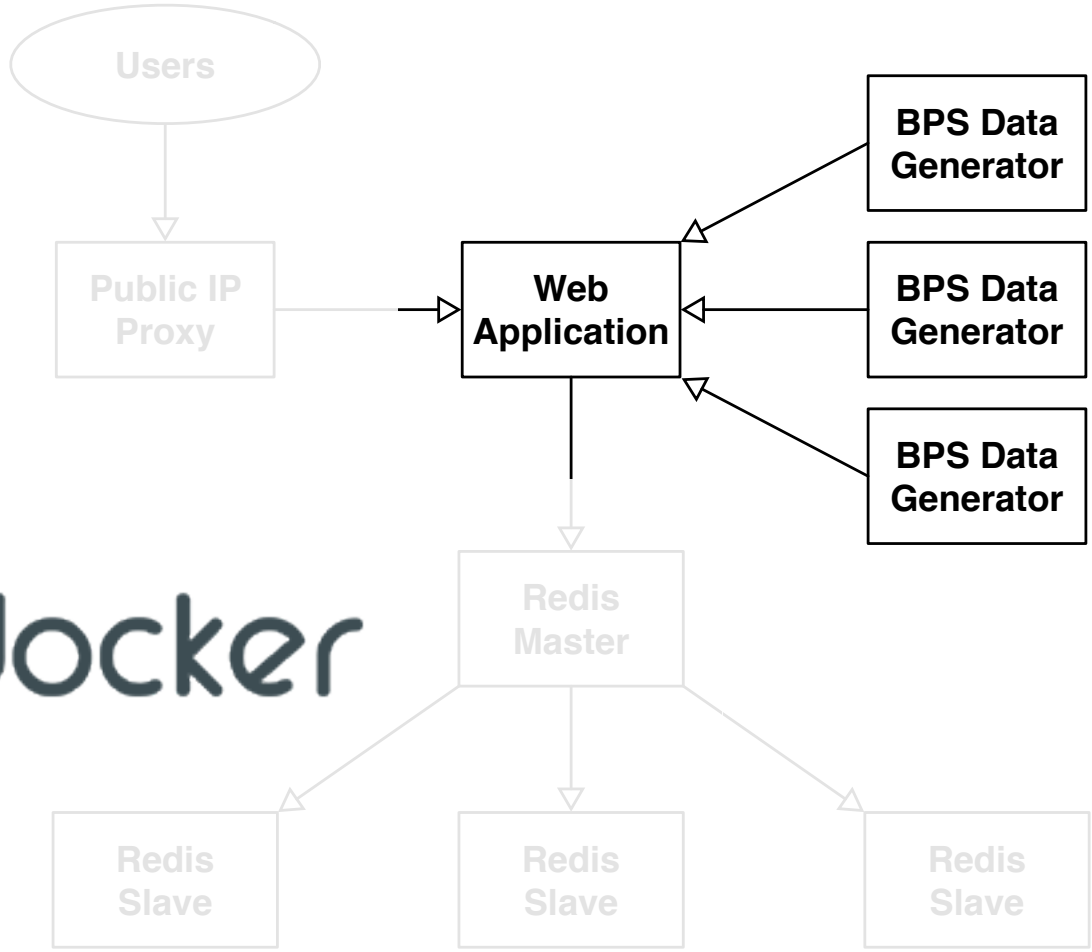
- Creating a new user
  - User Gluster permissions incorrect
- Cluster upgrade
  - Spark upgrade didn't take because of issue with Ansible role configuration
  - Wiped out our spark.conf – master / mesos settings wrong
- Gluster moint points disappeared on reboot
  - Not set in fstab

k8petstore


Users
Public IP Proxy
Web Application
BPS Data Generator
BPS Data Generator
BPS Data Generator
Redis Master
Redis Slave
Redis Slave
Redis Slave

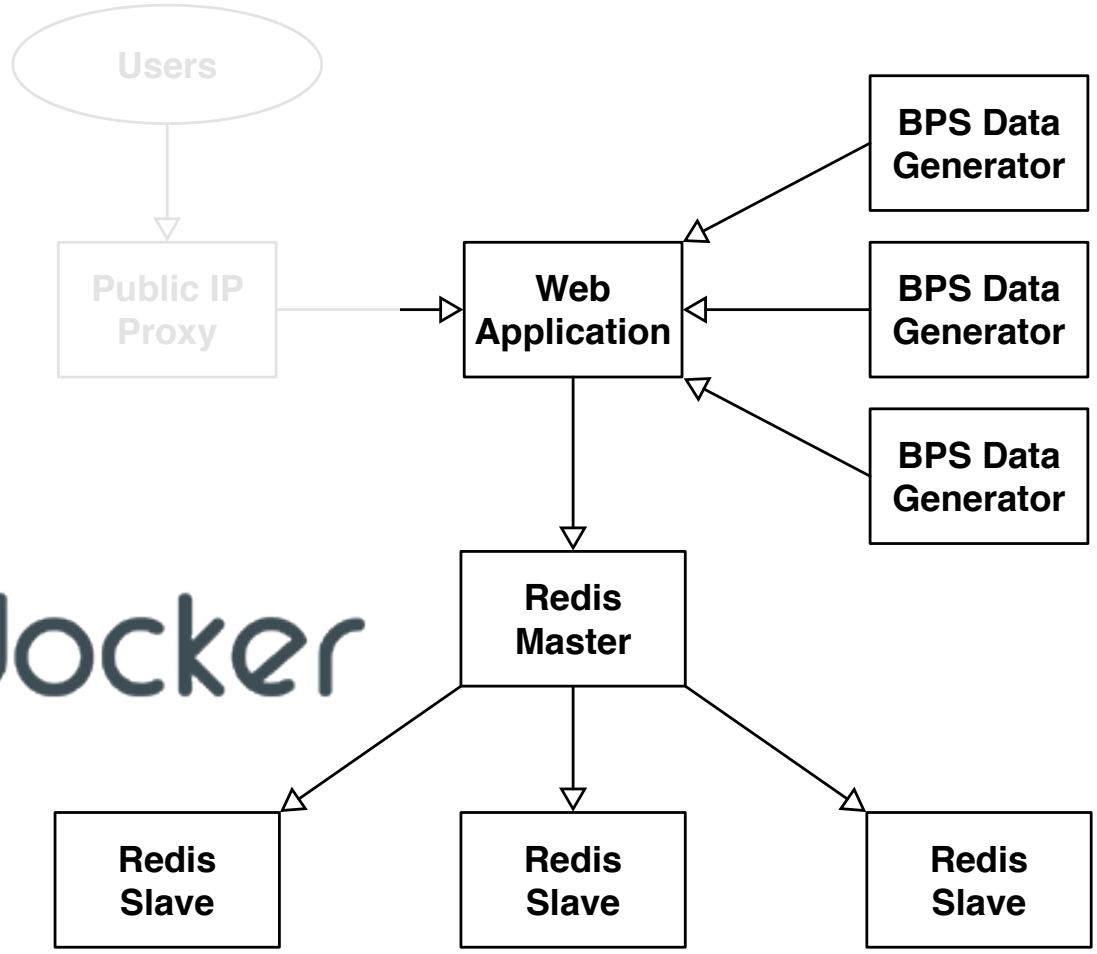

docker

k8petstore



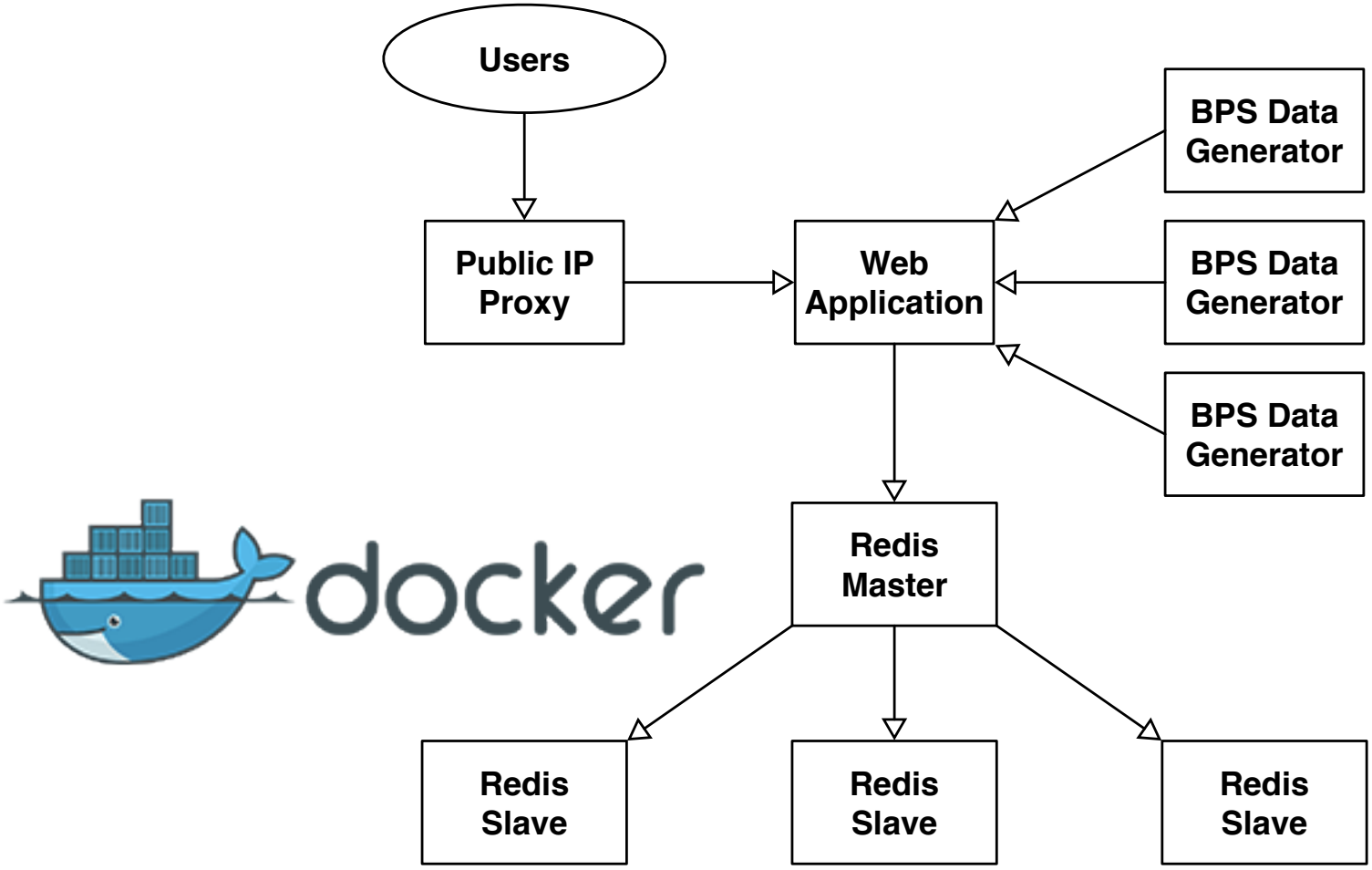
redhat

k8petstore


Users
Public IP Proxy
Web Application
BPS Data Generator
BPS Data Generator
BPS Data Generator
Redis Master
Redis Slave
Redis Slave
Redis Slave


docker

k8petstore


Users
Public IP Proxy
Web Application
BPS Data Generator
BPS Data Generator
BPS Data Generator
Redis Master
Redis Slave
Redis Slave
Redis Slave

k8petstore

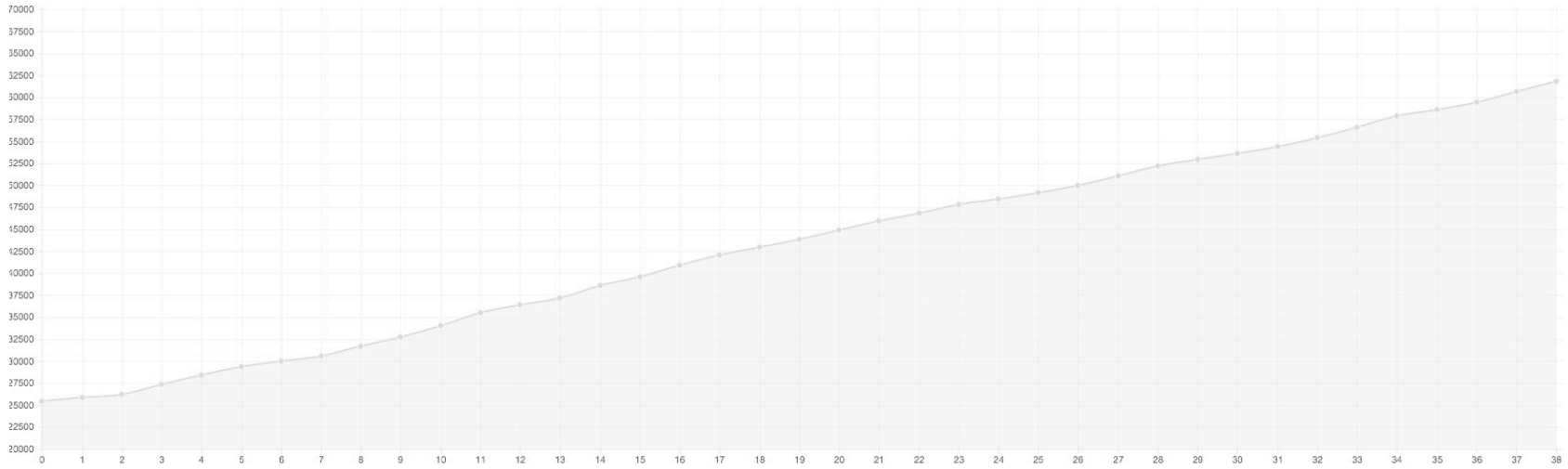
CURRENT TIME : 1426084741205

TOTAL entries : 61811
transaction delta 1164

0 _cost"]}],"customer":{"name":{"first":"Tymera","second":"Blakden"},"location":{"zipcode":"33168","coordinates":{"first":25.892185,"second":-80.21032},"city":"Miami","medianHouseholdIncome":43555.0,"po
1 _cost"]}],"customer":{"name":{"first":"Heavynn","second":"Goeff"},"location":{"zipcode":"11223","coordinates":{"first":40.598142,"second":-73.97229},"city":"Brooklyn","medianHouseholdIncome":40960.0,"
2 t_cost"]},{"fieldNames":["category","brand","size","per_unit_cost"]},{"fieldNames":["category","brand","flavor","size","per_unit_cost"]},{"fieldNames":["category","brand","color","size","per_unit_cost
3 t_cost"]}],"customer":{"name":{"first":"Heavynn","second":"Goeff"},"location":{"zipcode":"11223","coordinates":{"first":40.598142,"second":-73.97229},"city":"Brooklyn","medianHouseholdIncome":40960.0,

k8-bps.

http://host17-rack10.scale.openstack.engineering.redhat.com:3000/
/env /info



redhat.

Use Cases

- Configuration
- Scalability
- Fault Tolerance

k8petstore

- OpenContrail networking solution demo[1]
- Kubernetes JuJu Charm documentation example[2]
- Kubernetes v1.0 launch talk at OSCON[3]

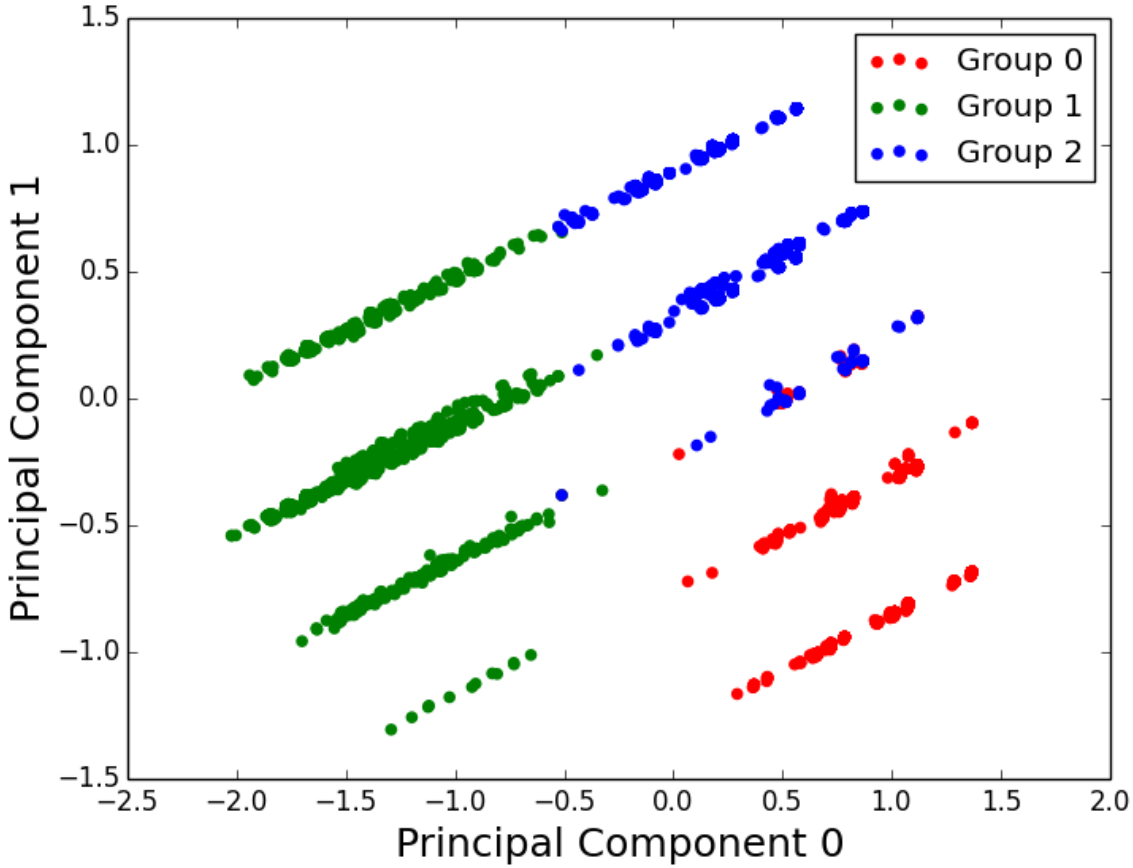
[1] - https://pedrormarques.wordpress.com/2015/04/24/kubernetes-and-opencontrail/

[2] - http://kubernetes.io/v1.0/docs/getting-started-guides/juju.html

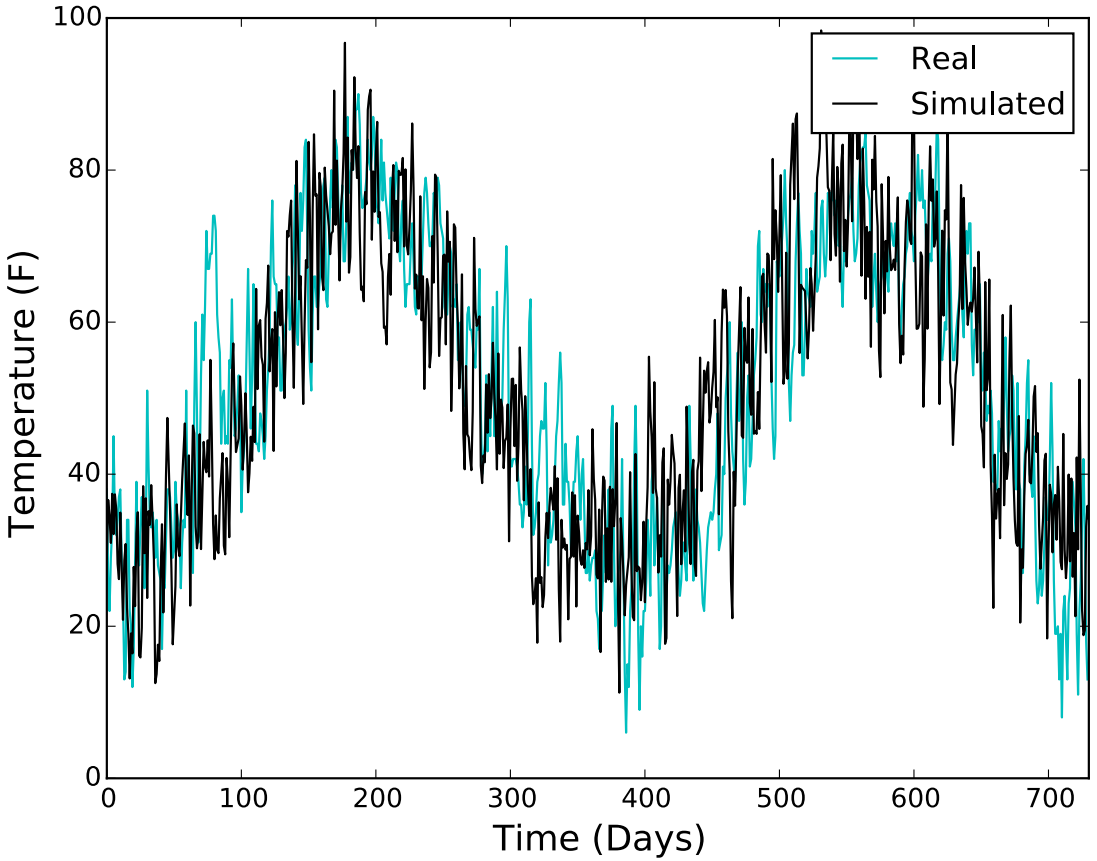
[3] - http://www.oscon.com/open-source-2015/public/schedule/detail/45281

APACHE BIGTOP DATA GENERATORS

BigPetStore

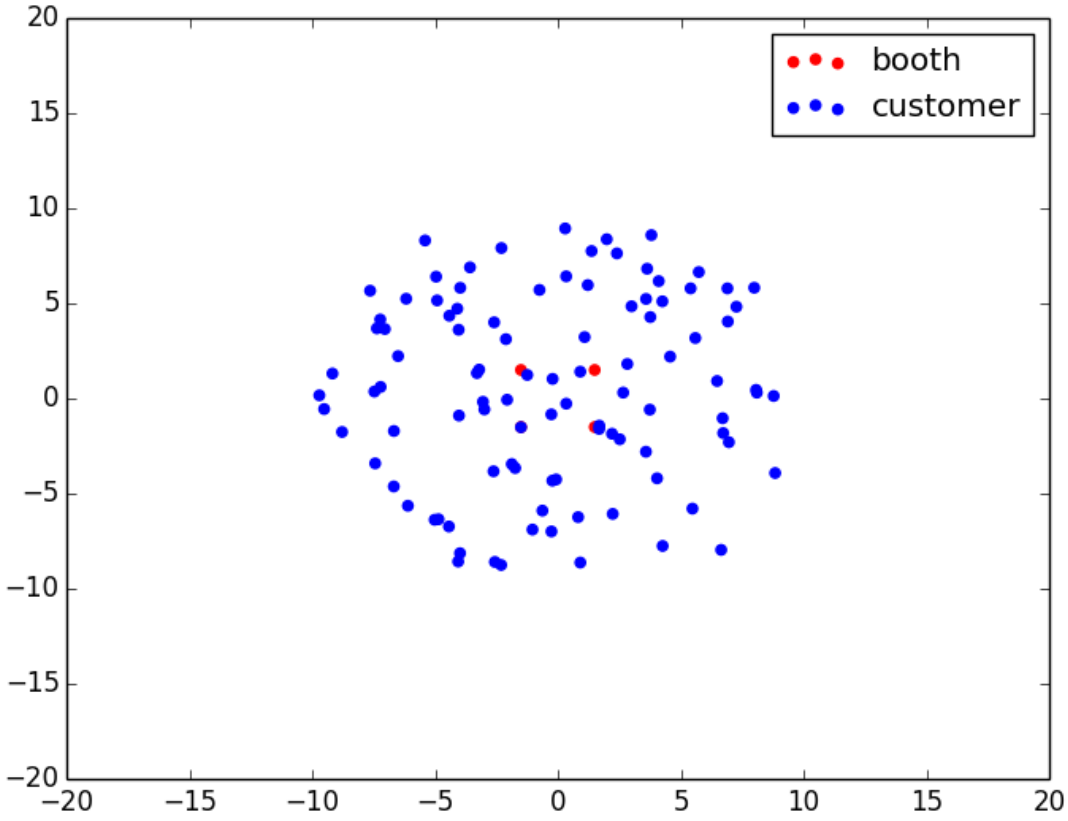


BigTop Weatherman



redhat.

BigTop Bazaar



Vision

- Encourage synthetic data generation for testing and realistic examples
- Serve as a resource for the larger Apache and open source communities
- Emphasis on
  - Flexibility
  - Scalability
  - Realism
- We look forward to collaborating and getting folks involved!

Resources

http://bigtop.apache.org/

http://github.com/apache/bigtop

http://rnowling.github.io/

Conclusion

- Synthetic data generators and blueprints are useful!
- Case studies:
  - Data Processing Pipelines
  - Cluster Deployment
  - Kubernetes
- BigPetStore and BigTop Data Generators efforts in Apache BigTop
- Open invitation to get involved and collaborate

QUESTIONS