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

- **Raw Daily Page Views**
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Accounts
  - Daily Activity

- **Raw Daily Page Views**
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Aggregate

- **Raw Daily Page Views**
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Cumulative Activity
Data Transformation and Summarization Pipeline

- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Aggregate
  - Cumulative Activity
  - Daily Activity

- Accounts

Spark
Data Transformation and Summarization Pipeline

- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Aggregate

- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Aggregate

- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
  - Summarize
  - Aggregate

Accounts → Daily Activity → Cumulative Activity
Data Transformation and Summarization Pipeline

- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate
- Raw Daily Page Views
  - Transform Raw Text
  - Parse
  - Clean & Validate

Accounts

Summarize

Aggregated

Cumulative Activity

Daily Activity
Data Transformation and Summarization Pipeline

- Raw Daily Page Views → Transform Raw Text → Parse → Clean & Validate → Summarize → Accounts
- Raw Daily Page Views → Transform Raw Text → Parse → Clean & Validate → Summarize → Daily Activity
- Raw Daily Page Views → Transform Raw Text → Parse → Clean & Validate → Summarize → Aggregate → Cumulative Activity
Data Transformation and Summarization Pipeline

- Raw Daily Page Views
  - Transform Raw Text
    - Parse
      - Clean & Validate
        - Summarize
          - Aggregate
            - Cumulative Activity
  - Transform Raw Text
    - Parse
      - Clean & Validate
        - Summarize
          - Daily Activity
- Raw Daily Page Views
  - Transform Raw Text
    - Parse
      - Clean & Validate
        - Summarize
          - Aggregate
            - Cumulative Activity
- Raw Daily Page Views
  - Transform Raw Text
    - Parse
      - Clean & Validate
        - Summarize
          - Aggregate
            - Cumulative Activity
Synthetic Data

• Sensitive Data
  – Real data on cluster for scalability testing and validation
  – Synthetic data for local development and testing

• Needed 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

Data Generator

- Raw Daily Page Views
- Accounts
- Expected Daily Activity
- Expected Cumulative Activity
- Transformation and Summarization Pipeline
  - Daily Activity
  - Cumulative Activity
- Validation Script
Validation with Synthetic Data

- Data Generator
- Raw Daily Page Views
- Accounts
  - Expected Daily Activity
  - Expected Cumulative Activity
- Transformation and Summarization Pipeline
- Daily Activity
- Cumulative Activity
- Validation Script
Validation with Synthetic Data

Data Generator

- Raw Daily Page Views
- Accounts
- Expected Daily Activity
- Expected Cumulative Activity

Transformation and Summarization Pipeline

- Daily Activity
- Cumulative Activity
- Validation Script
Validation with Synthetic Data

Data Generator

Raw Daily Page Views

Accounts

Transformation and Summarization Pipeline

Expected Daily Activity

Expected Cumulative Activity

Daily Activity

Cumulative Activity

Validation Script
Validation with Synthetic Data

Data Generator

Raw Daily Page Views

Accounts

Transformation and Summarization Pipeline

Expected Daily Activity

Cumulative Activity

Expected Cumulative Activity

Validation Script

15
Validation with Synthetic Data

Data Generator

Raw Daily Page Views

Accounts

Transformation and Summarization Pipeline

Expected Daily Activity

Daily Activity

Cumulative Activity

Expected Cumulative Activity

Validation Script
Issues Tackled

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

BPS Data Generator

ETL

Sales Statistics

Item Recommender
BigPetStore

HCFS

Core (RDDs)

BPS Data Generator

ETL

Sales Statistics

Item Recommender
BigPetStore

HCFS

Spark SQL

BPS Data Generator

ETL

Sales Statistics

Core (RDDs)

Item Recommender
BigPetStore

Spark SQL

HCFS

Core (RDDs)

MLLib

ETL

BPS Data Generator

Sales Statistics

Item Recommender
Team Cluster

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

Redis Master

BPS Data Generator

BPS Data Generator

BPS Data Generator

Redis Slave

Redis Slave

Redis Slave
Use Cases

• Configuration
• Scalability
• Fault Tolerance
k8petstore

• OpenContrail networking solution demo¹
• Kubernetes JuJu Charm documentation example²
• Kubernetes v1.0 launch talk at OSCON³

APACHE BIGTOP DATA GENERATORS
BigTop Weatherman

![Graph showing temperature over time with real and simulated data.](image-url)
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!
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
Resources

http://bigtop.apache.org/

http://github.com/apache/bigtop

http://rnowling.github.io/
QUESTIONS