Hadoop Inspector

The Problem

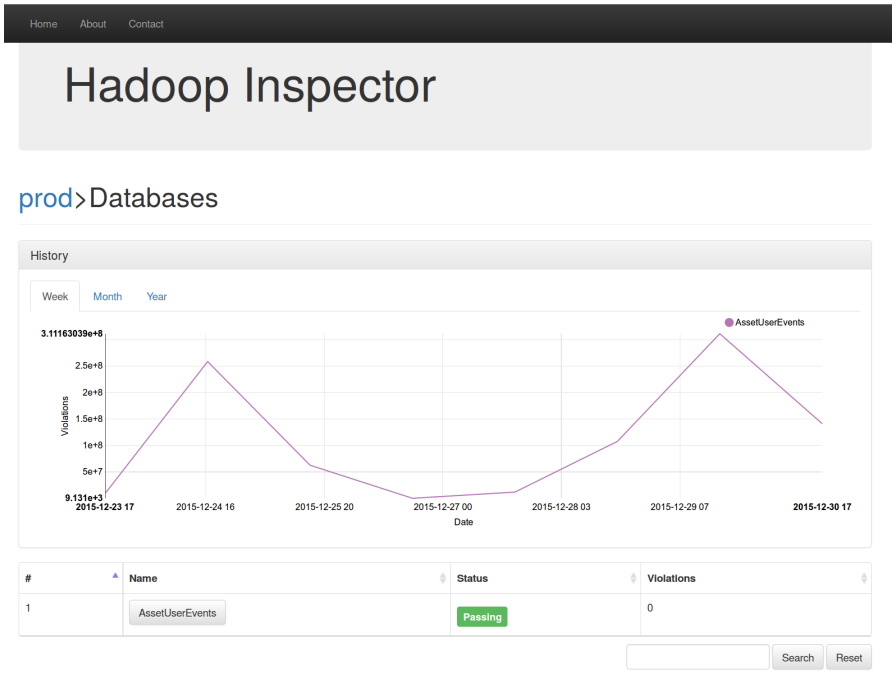
- Data Science has a low-tolerance for data quality problems
- Data Science has a high frequency of data quality problems
- Hadoop administration continues to be difficult, expensive, and error prone

The Consequences

- Loss of credibility
- Loss of algorithmic accuracy
- Loss of data scientist and analyst productivity
- Irreproducible findings
- Project failure or cancellation

The Solution: Hadoop Inspector

- Low-barrier-entry to writing tests
- Tests can be written in shell scripts, python, ruby, java
- Tests can include SQL, MapReduce, Spark
- Tests can integrate with ETL logs, external systems



Sample Check Types

- **Data Quality Checks:**
  - check uniqueness
  - check foreign key reference
  - check case
  - check min and max value
  - check min and max length
  - check accessibility
  - check type
  - check required fields
  - check unknown value
  - check enumerated value
  - check `start_timestamp` < `stop_timestamp`
- **Data Consistency Checks**
  - check base vs summary table
  - check peer tables
  - check target vs source
- **Data Management Checks:**
  - check statistics age
  - check retention age
  - check access
  - check blocksize
  - check file formats

Architecture

- **Reusable and Flexible Plug-in Architecture**
  - Phase One: Universal Plug-ins
    - checks are written in any language
    - runner passes run-time info to checks via env vars
    - checks pass results back to runner via stdout
  - Phase Two: SQL Plug-ins
    - checks are written as SQL
    - checks are a template that is filled-in by runner
  - Phase Three: Native Plug-ins
    - checks are written as Python modules
    - checks inherit from one another
    - runner imports plug-ins, has tight exception handling, logging, etc
- **Check Results Database to support historical forensics and data annotation.**
- **Both rule and profiling checks**
- **Version-control-compatible checks and registry - allowing test code to be managed with DDL.**

Licensing

Roadmap

- **2015-10-01: Initial Release**
  - Runner: initial release
  - Web Server: initial release
- **2015-11-01: Solidification**
- **2015-12-01: Functionality**
  - Support for native-SQL checks
- **Future**

Contact Information:
Will Farmer, Ken Farmer
willzfarmer@gmail.com
www.will-farmer.com
kenfar@gmail.com
www.linkedin.com/in/kenfar