

Avant Internship Review

Ben Wang

Avant

July 31, 2017

Data Quality Framework

Project 1: Handling NaNs

Problem

- ▶ Tradelines hourly check reported false positive alters
- ▶ NaNs in the report tables

Root Cause

- ▶ Merged two tables based on a categorical key
- ▶ Used NULL values when right table didn't contain a key

Solution

- ▶ Added a NaN-handling step
- ▶ replaced all NaNs with 0

Data Quality Framework

Project 2: Created Data Quality Checks for Credit Decision and Product Decision

My work

- ▶ Wrote SQL queries to select inputs and calculate summary statistics
- ▶ Merged the queries into DQF to create garden jobs

Difficulty

- ▶ All inputs nested in to one column as a String

Solution

- ▶ Used *plit_part* and *substring* to phrase inputs into individual into different variables

Direct Mail

Project 3: Fixed garden jobs that create suppression files

Problem

- ▶ Decline garden job could not finish

Root Cause

- ▶ The job's query was interrupted by followers

Solution

- ▶ Read the original query and document the logical structure in Confluence
- ▶ Translate query to Hive SQL
- ▶ Validate results of PostgreSQL and Hive query

- ▶ Your introduction goes here!
- ▶ Use `itemize` to organize your main points.

Examples

Some examples of commonly used commands and features are included, to help you get started.

Tables and Figures

- ▶ Use `tabular` for basic tables — see Table 1, for example.
- ▶ You can upload a figure (JPEG, PNG or PDF) using the files menu.
- ▶ To include it in your document, use the `includegraphics` command (see the comment below in the source code).

Item	Quantity
Widgets	42
Gadgets	13

Table 1: An example table.

Readable Mathematics

Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_i^n X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.