Final Project M2

Monday, March 27, 2017

Agenda

- Reading Quiz
- Review M2 Requirements
- M2 Support Session

Q1: Redshift is optimized for very fast execution of complex analytic queries against very large data sets.

a) True

b) False

Q2: What are the goals of a data distribution strategy?

a) To distribute the workload uniformly among the nodes in the cluster

b) To minimize data movement during query execution

c) Both are goals

d) Neither are goals

Q3: Which of the following is false?

a) **Even** Distribution distributes records in a round-robin style

b) **Uneven** Distribution distributes by record size

c) **Key** Distribution tries to place records with matching values on the same node slice

d) **All** Distribution puts a copy of the table on every node

Q4: What is the purpose of a leader node?

a) To protect other nodes from malware

b) To perform query tasks when other nodes are unavailable

c) To take credit for all the work done by other nodes

d) To manage the distribution of data and query processing tasks to the compute nodes

Q5: The COPY command can apply automatic compression during the load process

a) True

b) False

Final Project Datasets

Basic Specs:

- Discog:	8 files, 480MB
- Million Song:	36 files, 9G
- Music Brainz:	78 files, 6G

Main Entities:

- Discog:	Artists, Releases, Labels, Genres
- Million Song:	Artists, Songs, Tracks
- Music Brainz:	Artists, Tracks, Releases, Places, Events, Labels

Common Attributes across Datasets:

- Artist name
- Track name / title
- Release name / title

Song versus track?

https://joebennett.net/2012/05/18/song-vs-track-the-picture-and-the-frame/