Advanced Data Management

Medha Atre

Office: KD-219
atrem@cse.iitk.ac.in

Sept 1, 2016

Announcements

- Assignment-2 is online, it is due on Sept 8 23:59 IST.
- Course project proposals due on Sept 12, 2016, 23:59 IST. A larger list of categorized papers will be uploaded by the end of this week to aid you in defining the course project. Students can choose a topic outside the list as well, as long as it fits within the broader course objective.
- Try your hands at the Hadoop framework, the next programming assignment will be on Hadoop.

Graphs over MapReduce

- Distribution strategies:
  - Vertex-based distribution
  - Locality aware distribution with methods like METIS
  - n-hop distribution
- Distribution often happens as a pre-processing activity and not during the query execution.
- Query planner/optimizer takes the distribution strategy and any indexes into consideration and builds a query plan.
- Each compute node executes that query plan independently.

Data distribution with MapReduce – 1

- Import the data file using standard HDFS *put* or *copyFromLocal* commands.
- Run an *auxiliary* map-reduce job to first fetch this data and redistribute it according to your demand.
- Mapper's output key-value pair what you want to hash the data on.
- Reducer is an identity function.

Query Processing

- In case of data distribution strategy 1, the output of the previous auxiliary MapReduce job is used as an input to the query execution MapReduce job.
- Joins can be map-side or reduce-side.
- In case of map-side joins, reducers are often identity functions.
- In case of reduce-side joins, mappers are often identity functions.
- Mappers (or reducers) will in turn use standard query processing techniques like sort-merge-joins or hash-joins etc to join over the input data *keys*.

Let us solve our example on board