Relational Databases and Hadoop (in progress)

Dr. Garrett Dancik

What is a relational database?

- A relational database stores data in tables (with rows and columns)
- Adding data, modifying data, or accessing data therefore requires specifying the relevant table(s), column(s), and row(s).
- SQL, or structured query language, is a language for storing, manipulating, and retrieving data in databases.
- MySQL (https://www.mysql.com/) is an open source relational database management system (RDMS) that will be used in this class

```
SELECT * FROM students;
```

students

ID	firstName	lastname	age
1	Mary	Smith	20
2	Joe	Jackson	19
3	Barbara	Jones	21
4	Kevin	Waters	20

```
SELECT firstname, lastname
FROM students
WHERE age > 20;
```

students

ID	***firstName***	***lastname***	age
1	***Mary***	***Smith***	21
2	***Joe***	***Jackson***	19
3	***Barbara***	***Jones***	21
4	***Kevin***	***Waters***	20

firstName	***lastname***
Mary	***Smith***
Barbara	***Jones***

Hadoop and relational databases

- How do we transfer data from a relational database to HDFS? (Answer: Sqoop)
- Can we use a structured query language to interact with data on HDFS?
  - The Apache Hive (https://hive.apache.org/) data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL. Hive uses an SQL-like language (HiveQL) that submits MapReduce jobs
  - Impala (https://impala.apache.org/) is a Massive Parallel Processing SQL query engine for processing huge volumes of data stored on a Hadoop cluster. Uses special processes (daemons) that run on the cluster.