What is Apache Drill?

- A SQL query engine for a variety of non-relational databases and data files
  - *Hive, MongoDB, HBase*   Not Cassandra
  - *Even flat JSON or Parquet files on HDFS, S3, Azure, Google cloud, local file system*

- Based on Google's Dremel

BigTable --> HBase

GFS --> HDFS

Dremel --> Drill

It's real SQL

- Not SQL-Like

- And it has a ODBC / JDBC driver so other tools can connect to it just like any relational database
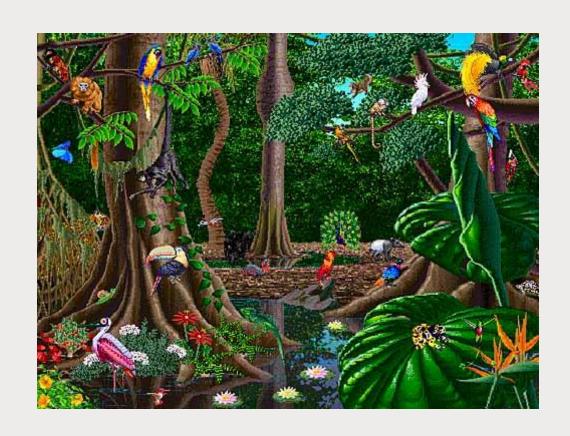
It's fast and pretty easy to set up.

- But remember, these are still <u>non-relational databases under the hood!</u>

- Allows SQL analysis of disparate data source without having to transform and load it first

  - *Internally data is represented as JSON and so has no fixed schema*

You can even do joins across different database technologies

- Or with flat JSON files that are just sitting around

Think of it as SQL for your entire ecosystem

Let's drill

- We'll import data into Hive and MongoDB

- Set up Drill on top of both

- And do some queries!