Data Pipelines on Apache Spark

@samelamin
Sam I Am

Who am I?
Why I Changed Careers?
WTF Is Big Data

Dan Ariely

January 6, 2013

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

2.5K Likes 124 Comments 1.2K Shares
The 3 Vs
ETL??
Database VS Data Warehouse

SAME DIFFERENCE

@samelamin
Datawarehouse vs Data Lake

"DATA LAKE"
Redshift Datawarehouse

Amazon Redshift works with your analysis tools

JDBC/ODBC

@samelamin
State Of Play

THOU SHALT NOT HARDCODE!
Slow Updates

DON'T MIND ME

JUST WAITING FOR MY ETL LOAD TO FINISH

@samelamin
Building A Data Pipelines

NOBODY BUILDS

DATA PIPELINES LIKE I DO

@samelamin
Data Formats

- CSV
  - Human readable
  - Common (e.g. dump from DB)
  - Slow read and write
  - Limited schema evaluation
  - Columnar file format
  - Good compression
  - Slow to write, fast to read
  - Support limited schema evolution
  - Widely used in big data (e.g. Spark, Impala)

- JSON
  - Human readable, simple
  - Support schema evaluation

- Parquet
  - Interoperability: Avro/Binary or Avro/Json
  - Very good compression support
  - Excellent schema evolution support
  - Widely used for serialization
  - Widely used as a serialization platform

- Avro
  - Interoperability: Avro/Binary or Avro/Json
  - Very good compression support
  - Excellent schema evolution support
  - Widely used for serialization
  - Widely used as a serialization platform
### Columnar Storage

#### Sales

<table>
<thead>
<tr>
<th>Product</th>
<th>Customer</th>
<th>Date</th>
<th>Sale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beer</td>
<td>Thomas</td>
<td>2011-11-25</td>
<td>2 GBP</td>
</tr>
<tr>
<td>Beer</td>
<td>Thomas</td>
<td>2011-11-25</td>
<td>2 GBP</td>
</tr>
<tr>
<td>Vodka</td>
<td>Thomas</td>
<td>2011-11-25</td>
<td>10 GBP</td>
</tr>
<tr>
<td>Whiskey</td>
<td>Christian</td>
<td>2011-11-25</td>
<td>5 GBP</td>
</tr>
<tr>
<td>Whiskey</td>
<td>Christian</td>
<td>2011-11-25</td>
<td>5 GBP</td>
</tr>
<tr>
<td>Vodka</td>
<td>Alexei</td>
<td>2011-11-25</td>
<td>10 GBP</td>
</tr>
<tr>
<td>Vodka</td>
<td>Alexei</td>
<td>2011-11-25</td>
<td>10 GBP</td>
</tr>
</tbody>
</table>

#### Product

<table>
<thead>
<tr>
<th>ID</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Beer</td>
</tr>
<tr>
<td>2</td>
<td>Beer</td>
</tr>
<tr>
<td>3</td>
<td>Vodka</td>
</tr>
<tr>
<td>4</td>
<td>Whiskey</td>
</tr>
<tr>
<td>5</td>
<td>Whiskey</td>
</tr>
<tr>
<td>6</td>
<td>Vodka</td>
</tr>
<tr>
<td>7</td>
<td>Vodka</td>
</tr>
</tbody>
</table>

#### Customer

<table>
<thead>
<tr>
<th>ID</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thomas</td>
</tr>
<tr>
<td>2</td>
<td>Thomas</td>
</tr>
<tr>
<td>3</td>
<td>Thomas</td>
</tr>
<tr>
<td>4</td>
<td>Christian</td>
</tr>
<tr>
<td>5</td>
<td>Christian</td>
</tr>
<tr>
<td>6</td>
<td>Alexei</td>
</tr>
<tr>
<td>7</td>
<td>Alexei</td>
</tr>
</tbody>
</table>
DAG Ting!
Data Lake Structure

AWS S3

DATA LAKE

Raw(CSV/JSON)

Raw(Parquet)

Modelled Data(Parquet)

DWH Mirror(Parquet)
Stateless Runs

IDEMPOTENT JOKES

ARE FUNNY NO MATTER HOW MANY TIMES YOU TELL THEM

@samelman
Handle Bad Data

NOT COOL BRO.
CLEAN YOUR DATA.
Schema Validation

SPARTANS DON'T LIKE BAD DATA!
Data warehouse On S3
Separate Storage Vs Compute
Beware corrections!

YO DAWG

I HEARD YOU LIKE CORECTIONS, I MEAN CORRECTIONS
Snapshot Vs Transactional

To ensure data integrity...

we'll need to roll this back
Three Levels of testing

HOW MANY?

THREE
Transformation Tests

FINALLY

THE TRANSFORMATION IS COMPLETE.

@samelamin
End To End Integration Tests (A/B)
Invariant Validation

THAT DOESN'T MAKE SENSE

BUT KEEP GOING

@samelamin
Backfills Are Crucial!

I HAVE A VERY IMPORTANT ANNOUNCEMENT

@samelamin
Beware The weasels!

I AM WEASEL.

CREATED BY DAVID FEISS
Schema Updates!

ONE DOES NOT SIMPLY
PROCESS DATA

@samelamin
Alerting!

I SEE EVERYTHING

SO CLEARLY NOW

@samelamin
Data Serving Layer

You want data?
Here's your data
ML Powering Recommendations

Recommendations.

Recommendations everywhere.