The Hadoop Stack, Part 2
Introduction to HBase

CSE 40822 – Cloud Computing – Spring 2016
Prof. Douglas Thain
University of Notre Dame
Three Case Studies

• Workflow: Pig Latin
  • A dataflow language and execution system that provides an SQL-like way of composing workflows of multiple Map-Reduce jobs.

• Storage: HBase
  • A NoSQL storage system that brings a higher degree of structure to the flat-file nature of HDFS.

• Execution: Spark
  • An in-memory data analysis system that can use Hadoop as a persistence layer, enabling algorithms that are not easily expressed in Map-Reduce.
References

• Fay Chang et al, Bigtable: A Distributed Storage System for Structured Data, OSDI 2006
  • http://research.google.com/archive/bigtable.html


• Apache HBase Documentation
  • http://hbase.apache.org
From HDFS to HBase

• HDFS provides us with a filesystem consisting of arbitrarily large files that can only be written once and are should be read sequentially, end to end. This works fine for sequential OLAP queries like Pig.

• OLTP workloads want to read and write individual cells in a large table. (e.g. update inventory and price as orders come in.)

• HBase implements OLTP interactions on top of HDFS by using additional storage and memory to organize the tables, and writing them back to HDFS as needed.

• HBase is an open source implementation of the BigTable design published by Google.
HBase Data Model

• A database consists of multiple tables.
• Each table consists of multiple rows, sorted by row key.
• Each row contains a row key and one or more column families.
• Each column family is defined when the table is created.
• Column families can contain multiple columns. (family:column)
• A cell is uniquely identified by (table, row, family:column).
• A cell contains an uninterpreted array of bytes and a timestamp.
## Data in Tabular Form

<table>
<thead>
<tr>
<th>Key</th>
<th>Name</th>
<th>Home</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Last</td>
<td>Phone</td>
</tr>
<tr>
<td>101</td>
<td>Florian</td>
<td>Krepsbach</td>
<td>555-1212</td>
</tr>
<tr>
<td>102</td>
<td>Marilyn</td>
<td>Tollerud</td>
<td>555-1213</td>
</tr>
<tr>
<td>103</td>
<td>Pastor</td>
<td>Inqvist</td>
<td></td>
</tr>
</tbody>
</table>
## Data in Tabular Form

<table>
<thead>
<tr>
<th>Key</th>
<th>Name</th>
<th>Home Phone</th>
<th>Email</th>
<th>Office Phone</th>
<th>Email</th>
<th>FacebookID</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Florian Garfield Krepsbach</td>
<td>555-1212</td>
<td><a href="mailto:florian@wobegon.org">florian@wobegon.org</a></td>
<td>666-1212</td>
<td><a href="mailto:fk@phc.com">fk@phc.com</a></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>Marilyn Tollerud</td>
<td>555-1213</td>
<td>666-1213</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>Pastor Inqvist</td>
<td>555-1214</td>
<td><a href="mailto:inqvist@wels.org">inqvist@wels.org</a></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

New columns can be added at runtime.

Column families cannot be added at runtime.
Don’t be fooled by the picture: Hbase is really a **sparse** table.
Nested Data Representation

Table People ( Name, Home, Office )
{
    101: {
        Timestamp: T403;
        Name: { First="Florian", Middle="Garfield", Last="Krepsbach" },
        Home: { Phone="555-1212", Email="florian@wobegon.org" },
        Office: { Phone="666-1212", Email="fk@phc.com" }
    },
    102: {
        Timestamp: T593;
        Name: { First="Marilyn", Last="Tollerud" },
        Home: { Phone="555-1213" },
        Office: { Phone="666-1213" }
    },
    ...
}
Nested Data Representation

GET People:101
101: {
    Timestamp: T403;
    Name: {First="Florian", Middle="Garfield", Last="Krepsbach"},
    Home: {Phone="555-1212", Email="florian@wobegon.org"},
    Office: {Phone="666-1212", Email="fk@phc.com"}
}

GET People:101:Name
People:101:Name: {First="Florian", Middle="Garfield", Last="Krepsbach"}

GET People:101:Name:First
People:101:Name:First="Florian"
Fundamental Operations

• CREATE table, families
• PUT table, rowid, family:column, value
• PUT table, rowid, whole-row
• GET table, rowid
• SCAN table (WITH filters)
• DROP table
Consistency Model

• Atomicity: Entire rows are updated atomically or not at all.
• Consistency:
  • A GET is guaranteed to return a complete row that existed at some point in the table’s history. (Check the timestamp to be sure!)
  • A SCAN must include all data written prior to the scan, and may include updates since it started.
• Isolation: Not guaranteed outside a single row.
• Durability: All successful writes have been made durable on disk.
The Implementation

• I’ll give you the idea in several steps:
  • Idea 1: Put an entire table in one file. (It doesn’t work.)
  • Idea 2: Log + one file. (Better, but doesn’t scale to large data.)
  • Idea 3: Log + one file per column family. (Getting better!)
  • Idea 4: Partition table into regions by key.

• Keep in mind that “one file” means one gigantic file stored in HDFS. We don’t have to worry about the details of how the file is split into blocks.
Idea 1: Put the Table in a Single File

File “People”

101: {Timestamp: T403; Name: {First="Florian", Middle="Garfield", Last="Krepsbach"}, Home: {Phone="555-1212", Email="florian@wobegon.org"}, Office: {Phone="666-1212", Email="fk@phc.com"}}, 102: {Timestamp: T593; Name: {First="Marilyn", Last="Tollerud"}, Home: {Phone="555-1213"}, Office: {Phone="666-1213"}}, . . .

• How do we do the following operations?
  • CREATE
  • DELETE
  • SCAN
  • GET
  • PUT
Variable-Length Data is **Fundamentally** Hard!

**SQL Table: People**

<table>
<thead>
<tr>
<th>ID</th>
<th>FirstName</th>
<th>LastName</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Florian</td>
<td>Krepsbach</td>
<td><strong>555-3434</strong></td>
</tr>
<tr>
<td>102</td>
<td>Marilyn</td>
<td>Tollerud</td>
<td>555-1213</td>
</tr>
<tr>
<td>103</td>
<td>Pastor</td>
<td>Ingvist</td>
<td>555-1214</td>
</tr>
</tbody>
</table>

```
UPDATE People SET Phone="555-3434" WHERE ID=403;
```

Each row is exactly 52 bytes long.
To move to the next row, just `fseek(file,+52);`
To get to Row 401, `fseek(file,401*52);`
Overwrite the data in place.

**HBase Table People**

```
PUT People, 403, Home:Phone, 555-3434
```

```
101: {Timestamp: T403;Name: {First="Florian", Middle="Garfield", Last="Krepsbach"},Home: {Phone="555-1212", Email="florian@wobegon.org"},Office: {Phone="666-1212", Email="fk@phc.com"}},102: {Timestamp: T593;Name: { First="Marilyn", Last="Tollerud"},Home: { Phone="555-1213" },Office: { Phone="666-1213" }}, . . .
```

Fun reading on the topic:
http://www.joelonsoftware.com/articles/fog0000000319.html
Idea 2: One Tablet + Transaction Log

Table for People


Transaction Log for Table People:

PUT 101: Office: Phone = “555-3434”
PUT 102: Home: Email = mt@yahoo.com
….

Memory Cache for Table People:

PUT People: 101: Office: Phone = “555-3434”

Changes are applied only to the log file. Then the resulting record is cached in memory. Reads must consult both memory and disk.

GET People: 101
GET People: 103
Idea 2 Requires Periodic Log Compression

Table for People on Disk: (Old)

101: {Timestamp: T403; Name: {First=“Florian”, Middle=“Garfield”, Last=“Krepsbach”}, Home: {Phone=“555-1212”, Email=“florian@wobegon.org”}, Office: {Phone=“666-1212”, Email=“fk@phc.com”}}, 102: {Timestamp: T593; Name: {First=“Marilyn”, Last=“Tollerud”}, Home: {Phone=“555-1213”}, Office: {Phone=“666-1213”}}, ...

Transaction Log for Table People:

PUT 101: Office: Phone = “555-3434”
PUT 102: Home: Email = mt@yahoo.com
....

Table for People on Disk: (New)

101: {Timestamp: T403; Name: {First=“Florian”, Middle=“Garfield”, Last=“Krepsbach”}, Home: {Phone=“555-1212”, Email=“florian@wobegon.org”}, Office: {Phone=“555-3434”, Email=“fk@phc.com”}}, 102: {Timestamp: T593; Name: {First=“Marilyn”, Last=“Tollerud”}, Home: {Phone=“555-1213”, Email=“my@yahoo.com”}}, ...

Write out a new copy of the table, with all of the changes applied. Delete the log and memory cache, and start over.
Idea 3: Partition by Column Family

Tablets for People on Disk: (Old)

- Data for Column Family Name
- Data for Column Family Home
- Data for Column Family Office

Transaction Log for Table People:

- PUT 101:Office:Phone = “555-3434”
- PUT 102:Home:Email = mt@yahoo.com
- ...

Tablets for People on Disk: (New)

- Data for Column Family Name
- Data for Column Family Home (Changed)
- Data for Column Family Office (Changed)

Write out a new copy of the tablet, with all of the changes applied. Delete the log and memory cache, and start over.
Idea 4: Split Into Regions

Region 1: Keys 100-200
Region 2: Keys 200-300
Region 3: Keys 300-400
Region 4: Keys 400-500

Where are the servers for table People?

Access data in tables

(Detail of One Region Server)

Tablet
Tablet
Tablet

Column Family Name
Column Family Home
Column Family Office
<table>
<thead>
<tr>
<th>Region</th>
<th>Keys</th>
<th>Column Family Name</th>
<th>Column Family Home</th>
<th>Column Family Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region 1</td>
<td>101-200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region 2</td>
<td>201-300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region 3</td>
<td>301-400</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table People
The consistency model is tightly coupled to the scalability of the implementation!
Consistency Model

• Atomicity: Entire rows are updated atomically or not at all.
• Consistency:
  • A GET is guaranteed to return a complete row that existed at some point in the table’s history. (Check the timestamp to be sure!)
  • A SCAN must include all data written prior to the scan, and may include updates since it started.
• Isolation: Not guaranteed outside a single row.
• Durability: All successful writes have been made durable on disk.
Questions for Discussion

• What are the tradeoffs between having a very tall vertical table versus having a very wide horizontal table?

• Suppose a table starts off small, and then a large number of rows are added to the table. What happens and what should the system do?

• Using the People example, come up with some examples of simple queries that are very fast, and some that are very slow. Is there anything that can be done about the slow queries?

• Would it be possible to have SCAN return a consistent snapshot of the system? How would that work?