Agenda

- Project overview
- Solr and SolrCloud
- Solr for indexing the events
- Hadoop
- Indexing using Hadoop and SolrCloud
- Web Interface
- Overall Architecture
- Screen Shots
Overview

- A tremendous amount ≈ 10TB of data is available about a variety of events crawled from the web.
- It is required to make this big data accessible and searchable conveniently through the web.
- ≈ 10TB of .warc.
- Use only HTML files.
Solr

- Solr is an open source enterprise search server based on the Lucene Java search library.

- Solr can be integrated with, among others...
  - PHP
  - Java
  - Python
  - JSON
SolrCloud

- What will happen if the server becomes full?
What is SolrCloud?

Shard & Replicate

- Scalability
- Fault Tolerance and Throughput

Shard 1
- Leader
- Replica
- Replica

Shard 2
- Leader
- Replica
- Replica
Schema

- `schema.xml` is usually the first file we configure when setting up a new Solr installation.

- The schema declares:
  - what kinds of fields there are
  - which field should be used as the unique/primary key
  - which fields are required
  - how to index and search each field
Schema (Cont.)

```xml
<fields>
  <field name="id" type="string" indexed="true" stored="true" required="true"/>
  <field name="name" type="textgen" indexed="true" stored="true"/>
...
</fields>

uniqueKey

```xml
  <uniqueKey>id</uniqueKey>
```

defaultSearchField

```xml
  <defaultSearchField>aggregate_text</defaultSearchField>
```

solrQueryParser

```xml
  <solrQueryParser defaultOperator="OR"/>
```
SolrCloud control

- solrctl instancedir --generate $HOME/solr_configs

- solrctl instancedir --create collection1 $HOME/solr_configs

- solrctl collection --create collection1 -s numOfShards

- http://128.173.49.32:8983/solr/#/~cloud
Event Fields

- We use the following fields
  - category: the event category or type
  - name: the event name
  - title: the file name
  - content: the file content
  - URL: the file path on the HDFS system
  - id: document ID
  - text: copy of the previous fields
Hadoop

- What is Hadoop?
- Features:-
  - Scalable
  - Economical
  - Efficient
  - Reliable
- Uses 2 main Services
  - HDFS
  - Map-Reduce
HDFS Architecture

Some Terminology

- **Job** – A “full program” - an execution of a Mapper and Reducer across a data set
- **Task** – An execution of a Mapper or a Reducer on a slice of data
- **Task Attempt** – A particular instance of an attempt to execute a task on a machine
MapReduce Overview

- User Program
- Master
- Workers

Input Data
- Split 0
- Split 1
- Split 2

Processes:
- Read
- Map
- Sort
- Reduce
- Write

Output Files:
- File 0
- File 1
MapReduce in Hadoop (1)
MapReduce in Hadoop (2)

Diagram showing the process of MapReduce in Hadoop, with input from HDFS, splitting, mapping, sorting, reducing, and output to HDFS with replication.
MapReduce in Hadoop (3)
Job Configuration Parameters

On cloudera

/user/lib/hadoop-*-mapreduce/conf/mapred-site.xml

```xml
<?xml version="1.0"?>
<configuration>

  <property>
    <name>mapred.reduce.tasks</name>
    <value>1</value>
    <description>The default number of reduce tasks per job</description>
  </property>

  <property>
    <name>io.sort.factor</name>
    <value>10</value>
    <description>Number of streams to merge at once while sorting</description>
  </property>

  <property>
    <name>io.sort.record.percent</name>
    <value>0.05</value>
    <description>Percentage of io.sort.mb dedicated to tracking record boundaries</description>
  </property>

</configuration>
```
Map TO Reduce

- Combiners
  - Often a map task will produce many pairs of the form \((k, v_1), (k, v_2), \ldots\) for the same key \(k\) (e.g., popular words in Word Count)
  - Can save network time by pre-aggregating at mapper
    - \(\text{combine}(k_1, \text{list}(v_1)) \rightarrow v_2\)
    - Usually same as reduce function

- Partition Function
  - For reduce, we need to ensure that records with the same intermediate key end up at the same worker
  - System uses a default partition function e.g., \(\text{hash(key)} \mod R\)
public class IndexerDriver extends Configured implements Tool {
    public static void main(String[] args) throws Exception {
        System.out.println("+---------------------------------------------------------------------+");
        System.out.println( "Version 3.0.0 ");
        System.out.println( "Connect to solr on port 8983 ");
        System.out.println("+---------------------------------------------------------------------");
        int exitCode = ToolRunner.run(new Configuration(),
            new IndexerDriver(), args);
        System.exit(exitCode);
    }

    @Override
    public int run(String[] args) throws Exception {
        JobConf conf = new JobConf(getConf(), IndexerDriver.class);
        conf.setJobName("Indexer");
        conf.setSpeculativeExecution(true);
        // Set Input and Output paths
        FileInputStream.setInputPaths(conf, new Path(args[0].toString()));
        FileOutputStream.setOutputPath(new Path(args[1].toString()));
        // Use TextInputFormat
        conf.setInputFormat(TextInputFormat.class);
        // Mapper has no output
        conf.setMapperClass(IndexMapper.class);
        conf.setMapOutputKeyClass(NullWritable.class);
        conf.setMapOutputValueClass(NullWritable.class);
        conf.setNumReduceTasks(0);
        JobClient.runJob(conf);
        return 0;
    }
}
@Override
public void map(LongWritable key, Text val, 
OutputCollector <NullWritable, NullWritable> output, 
Reporter reporter) throws IOException {

// st = new StringTokenizer(val.tostring());

txt="";

txt=val.toString();
    thisDoc = new SolrInputDocument();
    thisDoc.addField("id", fileName + key.toString()+key.toString());
    // Name of the file is the name of the event
    thisDoc.addField("name", eventname);
    // title is the file name
    thisDoc.addField("title", fileName);
    // file path is the file on URL on the HDFS
    thisDoc.addField("url", fileName+path);
    // category or the event calculated from the file path
    thisDoc.addField("category", cat);
    // content of the file
    thisDoc.addField("content", txt);
    
    try {
        server.add(thisDoc);
    } catch (SolrServerException e) {
        e.printStackTrace();
    } catch (IOException e) {
        e.printStackTrace();
    }
Solr REST API

- Solr is accessible through HTTP requests by using Solr’s REST API.
- Hard to create complex queries


- Results are returned as strings which requires some form of parsing.
["testovarHeaders": {
  "status": 0,
  "QTime": 5,
  "params": {
    "indent": "true",
    "q": "sisters",
    "hl": "false",
    "hl.simple.pre": "",
    "hl.simple.post": "",
    "wt": "json",
    "hl": "true"
  }
},
"response": {
  "numFound": 1,
  "start": 0,
  "docs": [
    {
      "id": "T21.txt00",
      "name": "Typhoon Haiyan",
      "text": [
        "Typhoon Haiyan",
        "T21.txt",
        "Friends, please join me today in not only praying for, but also actively supporting our brothers and sisters in the Philippines who are suffering in the wake of Typhoon H"
      ],
      "title": {
        "T21.txt"
      },
      "category": "Typhoon",
      "content": [
        "Friends, please join me today in not only praying for, but also actively supporting our brothers and sisters in the Philippines who are suffering in the wake of Typhoon H"
      ],
      ",version": 1460231380320560000
    }
  ],
  "highlighting": {
    "T21.txt00": {
      "text": [
        " and sisters in the Philippines who are suffering in the wake of Typhoon Haiyan"
      ]
    }
  }
}
Solarium

- Solarium is a PHP client for Solr to allow easy communication between PHP programs and the Solr server containing the indexed data.

- Solarium provides an Object Oriented interface to Solr which makes it easier for developers than the Solr’s REST API.

- Current version 3.2.0.
Why Solarium

- Solarium makes it easier for creating queries.
- Object Oriented interface rather than URL REST interface.

```php
$client = new Solarium\Client($config);
$query = $client->createSelect();
$query->setFields(array('id', 'url'));
$query->setStart($start)->setRows(10);

// get highlighting component and apply settings
$hl = $query->getHighlighting();
$hl->setFields('content');
$hl->setSimplePrefix('<b><span style="color: red;">');
$hl->setSimplePostfix('</b></span>');
$query->setQuery($keyword);`
• Solarium makes it easier for getting results
• Rather than parsing JSON or XML strings results are returned as a PHP associative arrays.

```php
$resultset = $client->execute($query);
$highlighting = $resultset->getHighlighting();
foreach ($resultset as $document) {
    foreach ($document as $field => $value) {
        echo "<b>$field</b>: $value<br>";
    }
}
$highlightedDoc = $highlighting->getResult($document->id);
if ($highlightedDoc) {
    foreach ($highlightedDoc as $field => $highlight) {
        echo "<b>$field</b>: " . implode(' ( ... ) ', $highlight) . '<br>';
    }
}
```
Interface Architecture

Web Interface (HTML) → Search requests (AJAX) → Server (PHP) → Solarium → Solr Server

Query → Response (JSON or XML) → MYSQL DB → Events Information

Results → Events Information → Events Information

Query → Response (Assoc. Array) → Events Information
<table>
<thead>
<tr>
<th>Event Categories</th>
<th>IDEAL Pages Web Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event Categories</strong></td>
<td></td>
</tr>
<tr>
<td>Accident</td>
<td></td>
</tr>
<tr>
<td>Bombing</td>
<td></td>
</tr>
<tr>
<td>Building Collapse</td>
<td></td>
</tr>
<tr>
<td>Disease outbreak</td>
<td></td>
</tr>
<tr>
<td>Earthquake</td>
<td></td>
</tr>
<tr>
<td>Explosion</td>
<td></td>
</tr>
<tr>
<td>Ferry Sank</td>
<td></td>
</tr>
<tr>
<td>Fire</td>
<td></td>
</tr>
<tr>
<td>Floods</td>
<td></td>
</tr>
<tr>
<td>Food Crisis</td>
<td></td>
</tr>
<tr>
<td>Gas Attack</td>
<td></td>
</tr>
<tr>
<td>Global Emergency</td>
<td></td>
</tr>
<tr>
<td>Hurricane</td>
<td></td>
</tr>
<tr>
<td>Plane Crash</td>
<td></td>
</tr>
<tr>
<td>Shooting</td>
<td></td>
</tr>
<tr>
<td>Snowstorm</td>
<td></td>
</tr>
</tbody>
</table>

**Number of Documents found**: 2

**Displaying Page Number**: 1

- **id**: B22.txt00
  - **url**: hdfs://localhost.localdomain:8020/user/cloudera/sample1/Bombing/Car_Bomb_Somali/B22.txt

- **id**: B21.txt00
  - **url**: hdfs://localhost.localdomain:8020/user/cloudera/sample1/Bombing/Car_Bomb_Somali/B21.txt
Screen Shots (Cont.)

IDEAL Pages Web Interface

Event Categories
- Accident
- Bombing
- Building Collapse
- Disease outbreak
- Earthquake
- Explosion
- Ferry Sank
- Fire
- Floods
- Food Crisis
- Gas Attack
- Global Emergency
- Hurricane
- Plane Crash
- Shooting
- Snowstorm

Number of Documents found: 2

Document Details

name: Car_Bomb_Somali
url: hdfs://localhost.localdomain:8020/user/cloudera/sample/Bombing/Car_Bomb_Somali/B22.txt
category: Bombing
content: A suicide car bomb exploded near Somalia’s Foreign Ministry on Tuesday, killing at least four people even as Somali and Kenyan leaders met and agreed to cooperate on military action against Islamist insurgents. The blast rattled control Mogadishu and killed at least three passers-by and the suicide bomber, said police official Ali Hassan. Six people were wounded and taken to the hospital, he said.

Close
Screen Shots (Cont.)

IDEAL Pages Web Interface

Event Categories
- Accident
- Bombing
- Building Collapse
- Disease outbreak
- Earthquake
- Explosion
- Ferry Sank
- Fire
- Floods
- Food Crisis
- Gas Attack
- Global Emergency
- Hurricane
- Plane Crash
- Shooting
- Snowstorm

Screen Shots (Cont.)

5/1/2014

Screen Shots (Cont.)

5/1/2014
Screen Shots (Cont.)

content: expectations, railway announces stock splitExplosives used to accelerate blaze at Gainford train derailment

id: 86983276f539c16e478387694bf5194
content: projects include implementing a new transportation management system, pushing the detection and derailment

id: c26175a2a544a6617fd92631b1ddd3b
content: of a train derailment on a bridge over the swollen Bow River in Calgary. Continue reading

id: 9ce855565a1606bb8d78738e25a75e5
content: A train derailment on the Bonnybrook rail bridge has forced the closure of Deerfoot trail in both

id: 908c214f1b8f130d03906c52bdf880ac0
content: bridge collapsed a year ago after a Union Pacific train derailment. The falling bridge crushed a vehicle

id: 091eb7575e05087f13bc2755b390
content: the public is safe after a freight train derailment east of Sudbury on Sunday. CP shares trickled

id: 70c77da4047f6488343f28d0e0a4d
content: River following a train derailment that threatened to spill cars carrying diesel-like petroleum diluent

id: d4ae83f9b19a9ae88f83f9eb2e23414
content: A train derailment on the Bonnybrook rail bridge has forced the closure of Deerfoot trail in

<< First | < Previous | Next > | Last >>
Document Details

name: Hurricane Sandy
url: hdfs://localhost.localdomain:8020/user/cloudera/sample_data/Typhoon/Hurricane_Sandy/T12.txt
category: Typhoon
content: The skies cleared but there is still the visual shock of the tragic loss of property and life now dealing with the aftermath of Hurricane Sandy. Here in the East Bay of Rhode Island the impact of Sandy was minimal compared to Newport County and other more southern parts of the state. I'm still in a visual shock over seeing how much destruction this storm caused in New York and New Jersey.

id: T11.txt
url: hdfs://localhost.localdomain:8020/user/cloudera/sample_data/Typhoon/Hurricane_Sandy/T12.txt

id: T12.txt
url: hdfs://localhost.localdomain:8020/user/cloudera/sample_data/Typhoon/Hurricane_Sandy/T12.txt

<< First | < Previous | Next >> | Last >>
Thank you

Mohammed Farghally
&
Ahmed Elbery