A quick trip from code profiling to file formats

Daniel Lanza, CERN Database group

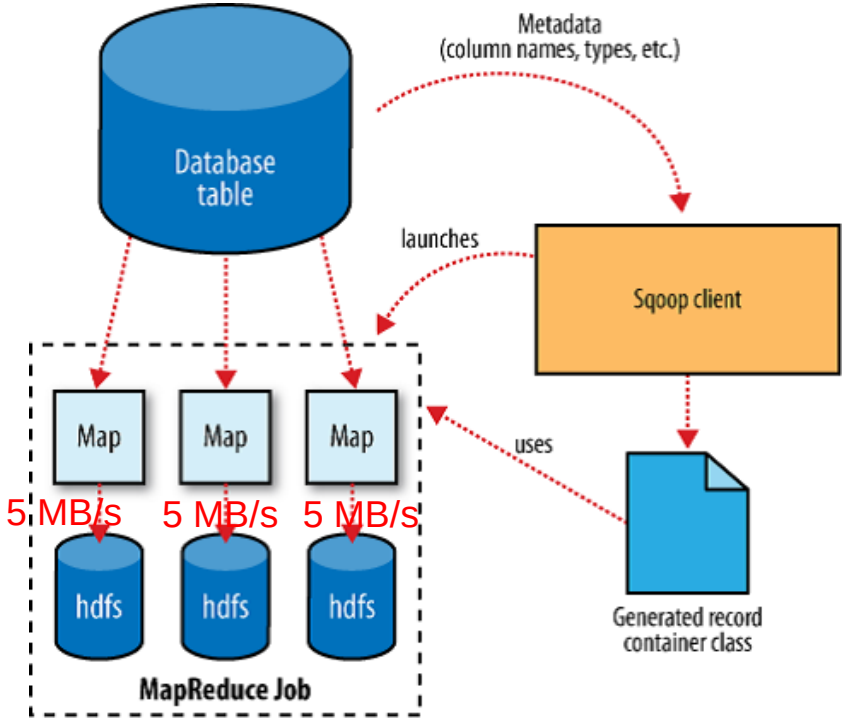
tCSC
27th May 2016

Agenda

- Importing data to Big Data solutions
  - Tool for importing (Java profiler for distributed applications)
  - Target file format (in numbers)
  - Compatile applications
  - Partitioning
  - Backups
  - Keep data up to date
  - …

Java profiler for distributed applications

Java profiler for distributed applications



CPU-bound!

Image source: T. White, The Definitive Hadoop Guide

Java profiler for distributed applications

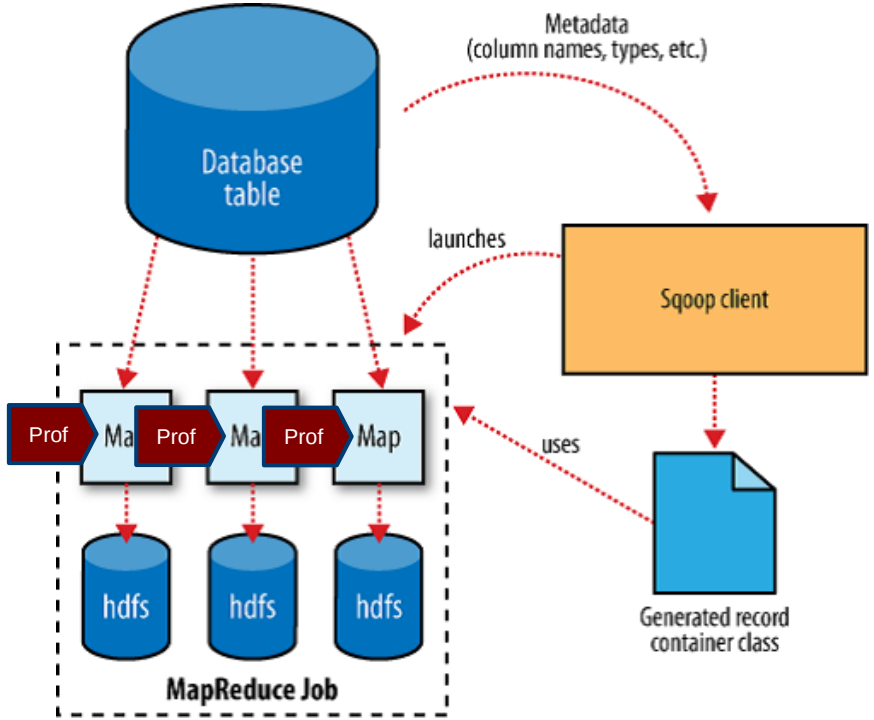
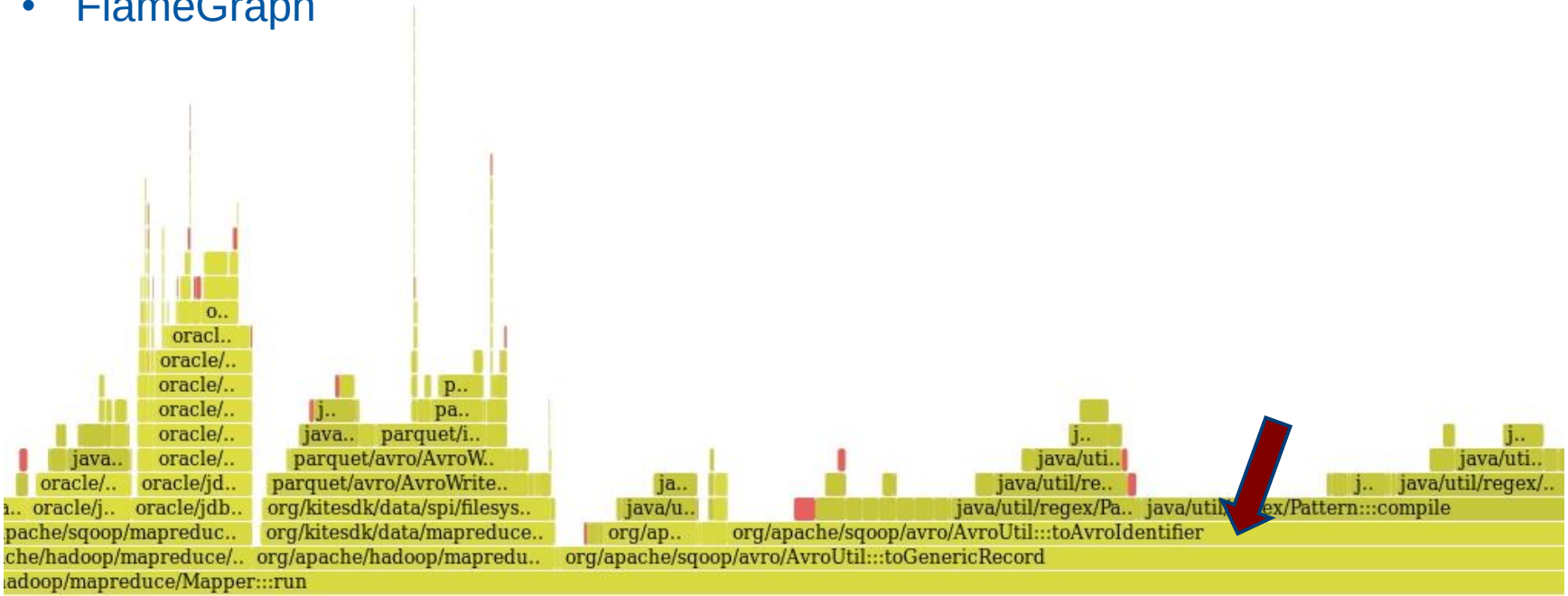


Image source: T. White, The Definitive Hadoop Guide

Java profiler for distributed applications

- FlameGraph



Java profiler for distributed applications

```java
/**
 * Format candidate to avro specifics
 */
public static String toAvroIdentifier(String candidate) {
  String formattedCandidate = candidate.replaceAll("\\W+", "_");
  if (formattedCandidate.substring(0,1).matches("[a-zA-Z_]")) {
    return formattedCandidate;
  } else {
    return "AVRO_" + formattedCandidate;
  }
}
```

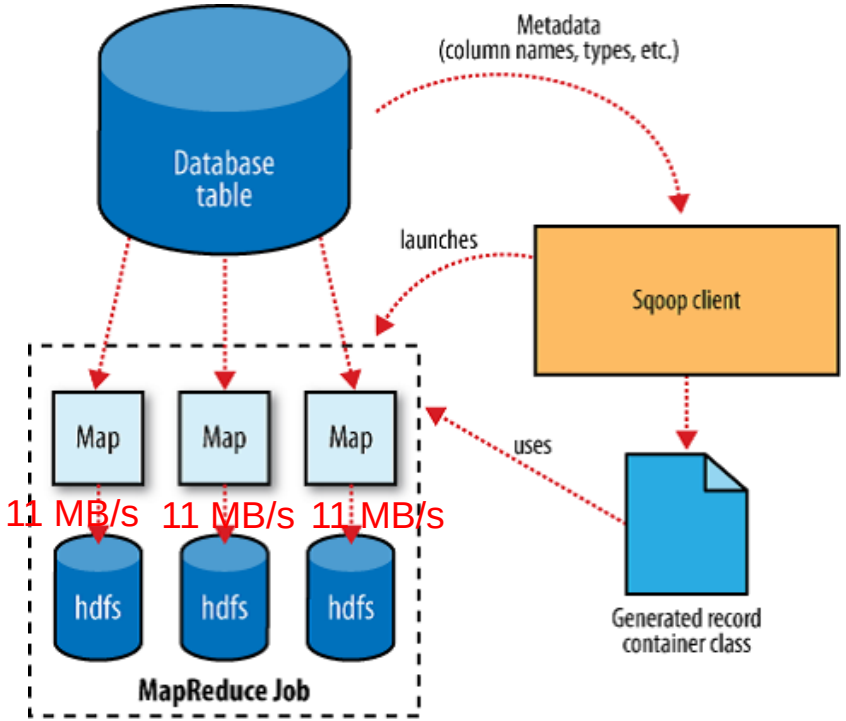
```java
/**
 * Format candidate to avro specifics
 */
public static String toAvroIdentifier(String candidate) {
  char[] data = candidate.toCharArray();
  boolean skip = false;
  int stringIndex = 0;

  for (char c:data) {
    if (Character.isLetterOrDigit(c) || c == '_') {
      data[stringIndex++] = c;
      skip = false;
    } else if(!skip) {
      data[stringIndex++] = '_';
      skip = true;
    }
  }

  char initial = data[0];
  if (Character.isLetter(initial) || initial == '_') {
    return new String(data, 0, stringIndex);
  } else {
    return "AVRO_".concat(new String(data, 0, stringIndex));
  }
}
```

500% faster!

Java profiler for distributed applications



Still CPU-bound

Image source: T. White, The Definitive Hadoop Guide

Java profiler for distributed applications

Sqoop / SQOOP-2906

Optimization of AvroUtil.toAvroIdentifier

Agile Board

Details

Type:	Improvement	Status:	RESOLVED
Priority:	Major	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	1.4.7
Component/s:	None		
Labels:	avro hadoop optimization		
Flags:	Patch		

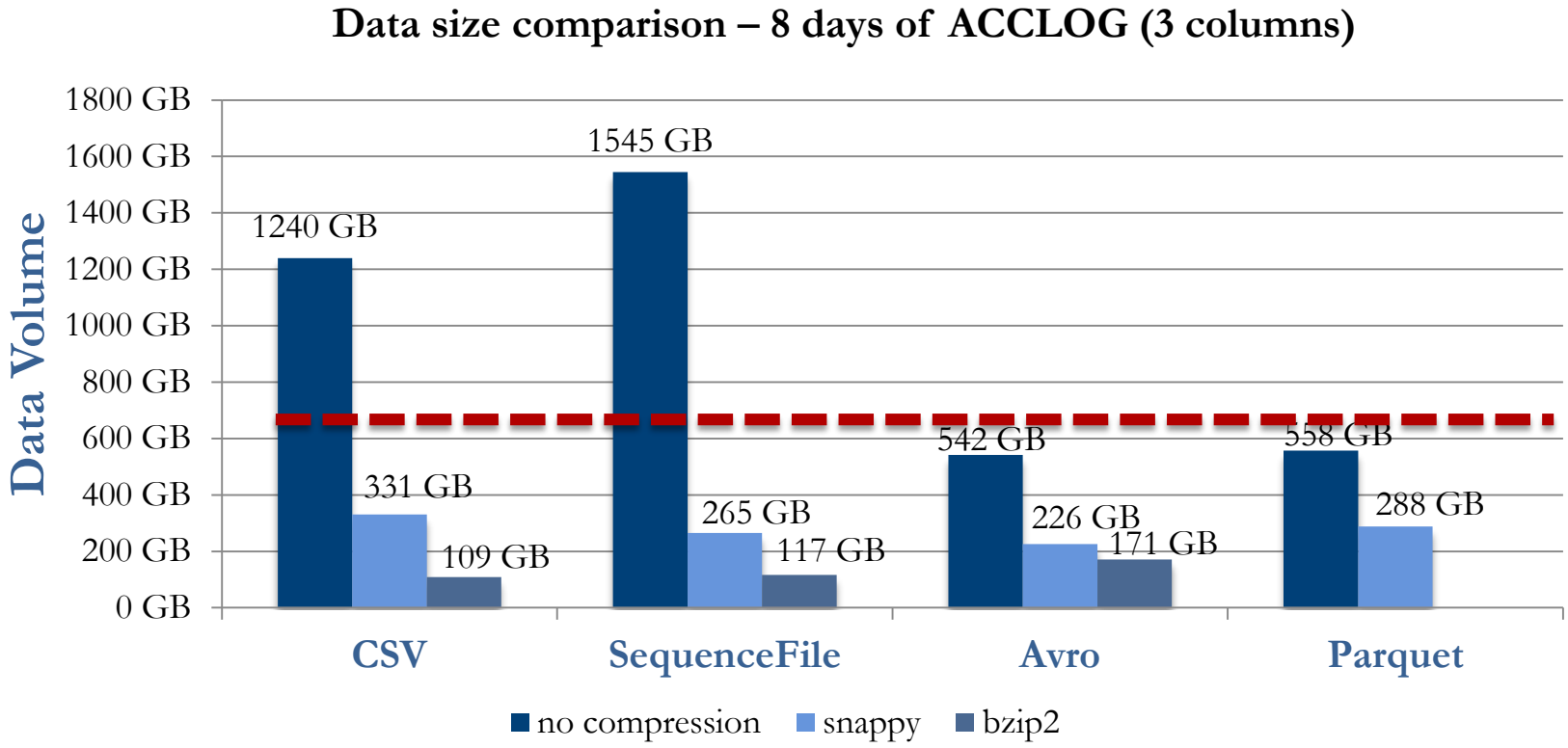
Java profiler for distributed applications

- CERN DB blog entry
  - http://db-blog.web.cern.ch/blog/prasanth-kothuri/2016-05-integrating-hadoop-and-elasticsearch-%E2%80%93-part-2-%E2%80%93-writing-and-querying

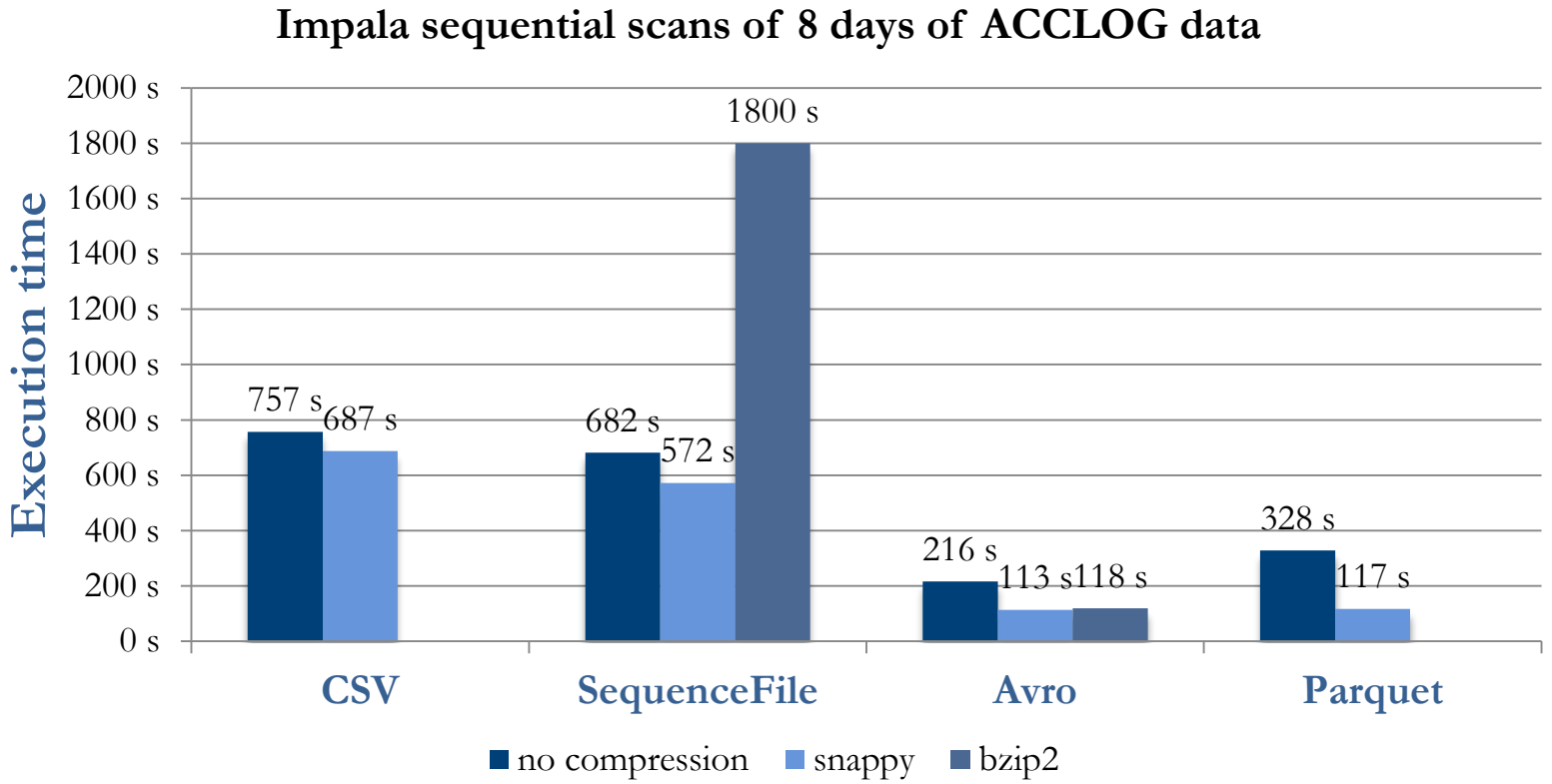
- Git repos
  - https://github.com/cerndb/Hadoop-Profiler
  - https://gitlab.cern.ch/jhermans/hprofiler

File formats in numbers

File formats in numbers

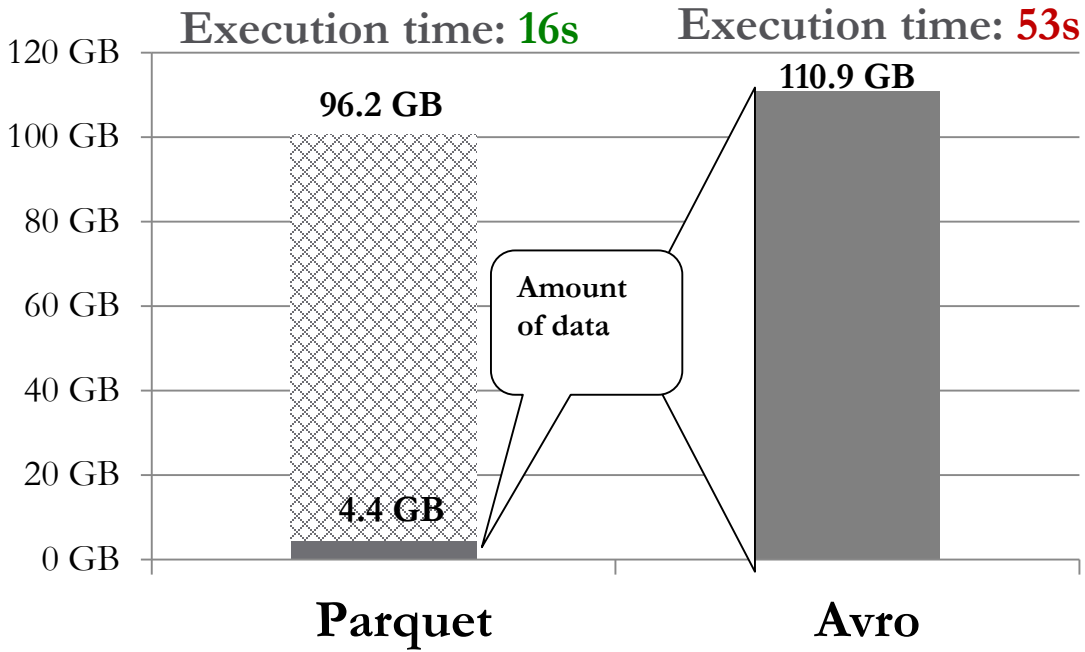


File formats in numbers



File formats in numbers

- Joining columns from different tables (1400 columns in total)



Messages to take away

- Profile your application if you find performance issues
- Choose carefully file format and compression

Questions / Feedback

Acknowledgements

- Zbigniew Baranowski
- Maciej Grzybek
- Kacper Surdy
- Joeri Hermans