That's not a data lake. This is a data lake.

By Eric Snyder

CHARIOT SOLUTIONS

What do we do with all of this data?

Data Warehouse, Data Lake, Data Ocean, Data Puddle

Easy Peasy Flink

AWS Data Lake

CHARIOT SOLUTIONS

Perspective

```
   average 2 sensors per hotel room
 x each sensor sends a message every 4 seconds (21,600 / day)
 x 256 bytes per event (sensor data, sensor id, account data)
```

> 5,000,000 hotel rooms in USA

```
0.5% of rooms contain sensors == 0.25 TiB / day
1.0% of rooms contain sensors == 0.5  TiB / day
5.0% of rooms contain sensors == 2.5  TiB / day
```

CHARIOT SOLUTIONS

Options

- **Time Series DB** - e.g. InfluxDB, Timescale DB.
  Ideal when time is primary.
- **Old Skool Relational** - e.g. postgresql
- **Relational Warehouse** - e.g. Amazon Redshift, Vertica, Snowflake
- **Data Lake**
  - persist as is or after on-the-fly transformation
  - “Just Land It”
  - often Hadoop-ish

CHARIOT SOLUTIONS

Fill That Lake

Short Term

Use Case: Chart the number of temperature changes in the last 5 minutes.

File/Object Formats: Row Based - avro, protobuf, capnproto, flatbuffers…

Long Term

Use Case: Find the top 10 hotels by water consumption per occupant in the last year.

File/Object Formats: Columnar - parquet, orc

CHARIOT SOLUTIONS

Why Avro?

- **Compact**: no wasted bytes, no row level field metadata. Compare to json. 1K of avro vs 1K of json.
- **Self-describing**: each file contains the schema definition used to write the file.
- **Schema evolution**: provide a reader schema that contains more or fewer fields than the writer schema.
- **Code generation (optional)**

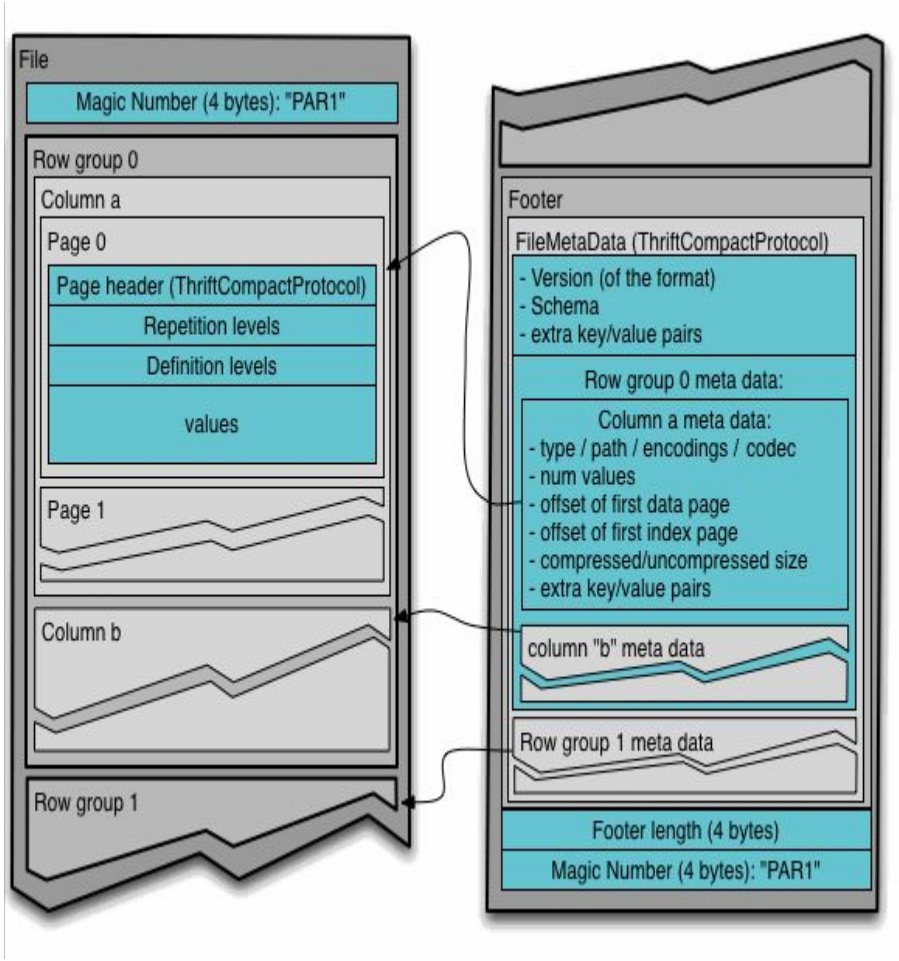
```
{
  "namespace": "example.avro",
  "type": "record",
  "name": "User",
  "fields": [
    {
      "name": "name",
      "type": "string"
    },
    {
      "name": "favorite_number",
      "type": ["int", "null"]
    },
    {
      "name": "favorite_color",
      "type": ["string", "null"]
    }
  ]
}
```

CHARIOT SOLUTIONS

Why Parquet?

Columnar storage: data from the same column is stored together, which facilitates:

- predicate pushdown
- column level compression/encoding
- column level metadata



CHARIOT SOLUTIONS

Apache Flink

- data processing and data-driven applications with *data streams* as the core building block
- low latency
- supports many data transformation and enrichment tasks
- event-time processing
- HA w. exactly-once state consistency
- an official AWS EMR application

```
// ...imports…

case class WordCount(word: String, count: Int)

object WindowWordCount {

  def main(args: Array[String]) {
    val env = StreamExecutionEnvironment
      .getExecutionEnvironment

    val text = env.socketTextStream(
      "localhost", 9999)

    val counts = text.flatMap {
       _.toLowerCase.split("\\W+")
    }
    .map { WordCount(_, 1) }
    .keyBy("word")
    .timeWindow(Time.seconds(5))
    .sum("count")

    counts.print

    env.execute("Window Stream WordCount")
  }

}
```

CHARIOT SOLUTIONS

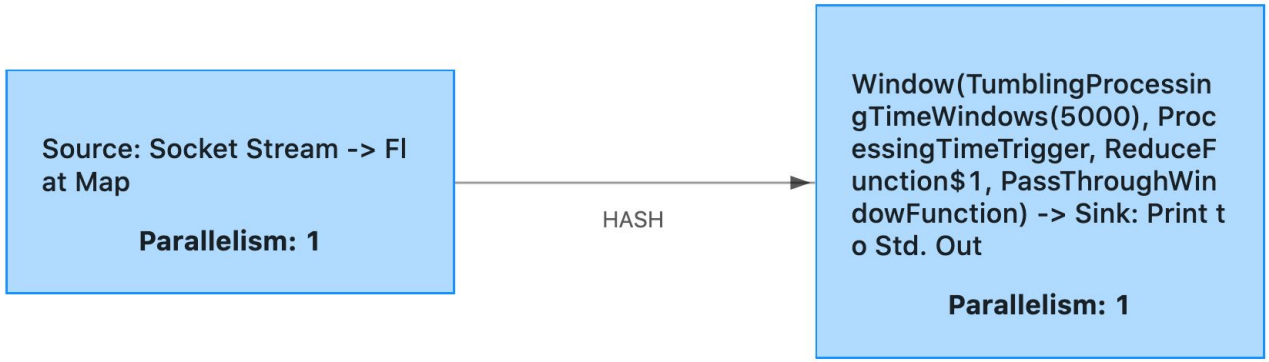
Socket Window WordCount

RUNNING | 2

ID: **3de928420bda94a63f06c36c5291f61a** | Start Time: **2019-10-15 22:22:16** | Duration: **2m 4s**

Cancel Job

Overview | Exceptions | TimeLine | Checkpoints | Configuration



Name	Status	Bytes Received	Records Received	Bytes Sent	Records Sent	Tasks
Window(TumblingProcessingTimeWindows(5000), ProcessingTi...	RUNNING	122 B	6	0 B	0	1
Source: Socket Stream -> Flat Map	RUNNING	0 B	0	0 B	6	1

Apache Beam

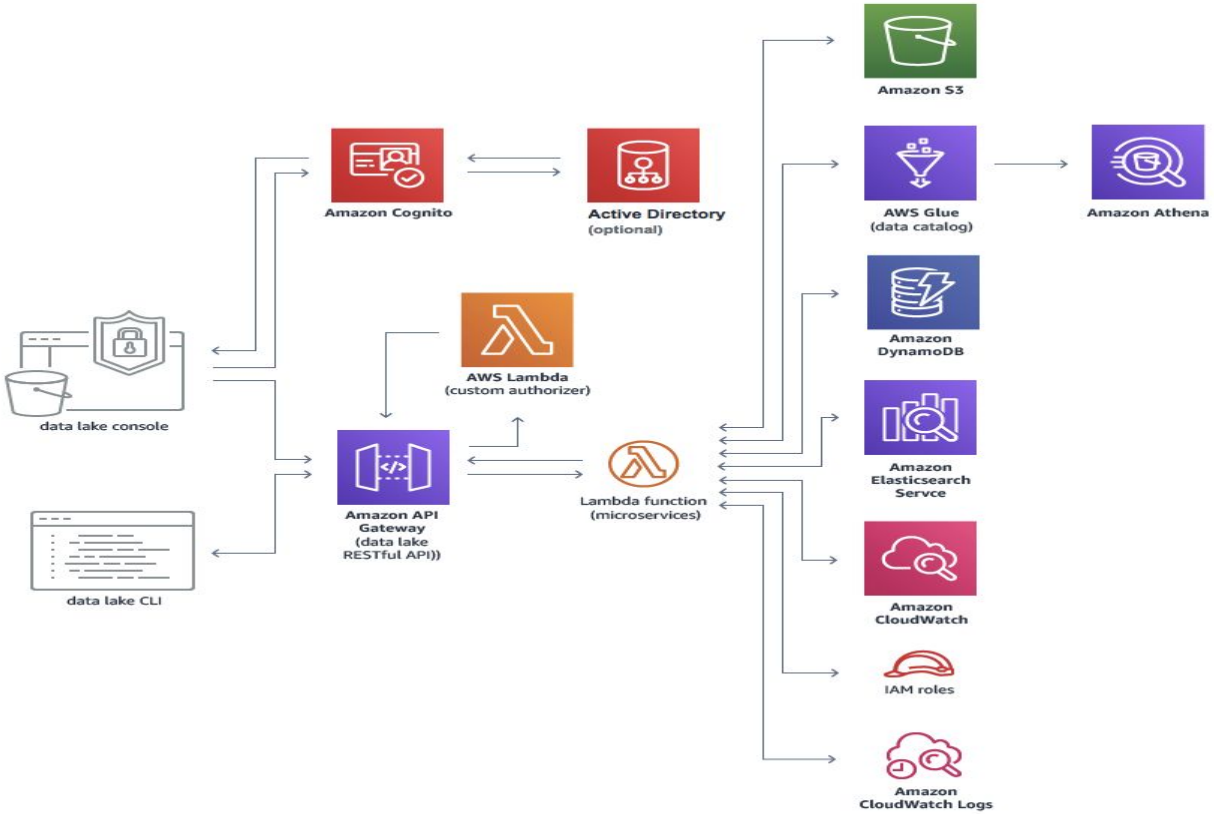
Unified Programming Model

- One programming model for batch and streaming.
- Execute pipelines on any supported or multiple execution environments (Spark, Flink, Google Cloud Dataflow, many others).
- Create and share common transformations and connectors.

CHARIOT SOLUTIONS

```
public class WindowedWordCount {
  static void runWindowedWordCount(Options options) {
    Pipeline pipeline = Pipeline.create(options);

    PCollection<String> input = pipeline
     .apply(
       TextIO.read().from(options.getInputFile()))
     .apply(
       ParDo.of(new AddTimestampFn(...)));

    PCollection<String> windowedWords = input
      .apply(
        Window.into(FixedWindows.of(
          Duration.standardMinutes(10))));

    PCollection<KV<String, Long>> wordCounts =
      windowedWords.apply(new WCount.CountWords());

    wordCounts
      .apply(
        MapElements.via(new WCount.FormatAsTextFn())
      )
      .apply(new WriteOneFilePerWindow(output, 8));

    PipelineResult result = pipeline.run();
    result.waitUntilFinish();
  }
}
```

AWS Data Lake



CHARIOT SOLUTIONS

References

Apache Flink

Apache Beam

AWS Data Lake

CHARIOT SOLUTIONS

Questions?

CHARIOT SOLUTIONS

Technology in the Service of Business.

Chariot Solutions is the Greater Philadelphia region's top IT consulting firm specializing in software development, systems integration, mobile application development and training.

Our team includes many of the top software architects in the area, with deep technical expertise, industry knowledge and a genuine passion for software development.

Visit us online at chariotsolutions.com.

CHARIOT SOLUTIONS