

CS 378 – Big Data Programming

Lecture 10

Data Organization Patterns

Review

- Assignment 4 – Avro Objects
- We'll look at implementation details of:
 - Mapper
 - Combiner
 - Should we use one? Can we use one?
 - Reducer
 - Avro generated Java code

AVRO Field Definitions

- Unions
 - With defaults
- Enumerations
 - In unions
 - With defaults

Design Pattern

- Structured to hierarchical design pattern
- Data sources linked by some foreign key
- Data is structured and row based
 - For example, from databases
- Data is semi-structured and event based
 - Web logs

Sessionizing Web Logs

- Create user sessions from web logs
- Represents all the actions by a user
- Allows later analysis to “replay” the user actions
- Collect measures and metrics about user behavior
 - Pages viewed, time on page, clicks
 - Path through the site, entry to the site (from a search engine?)

Sessionizing Web Logs

- To start (this or any “big data” application)
- We need to understand the data
 - Fields, values
 - Data size
- We need to define our goal
 - What do we want to end up with

Web Logs

- Let's look at some data
- Logs saved in database
 - Log entries already have structure
 - Tab separated values
 - Easily parsed (lots of work has been done for us)

Web Logs

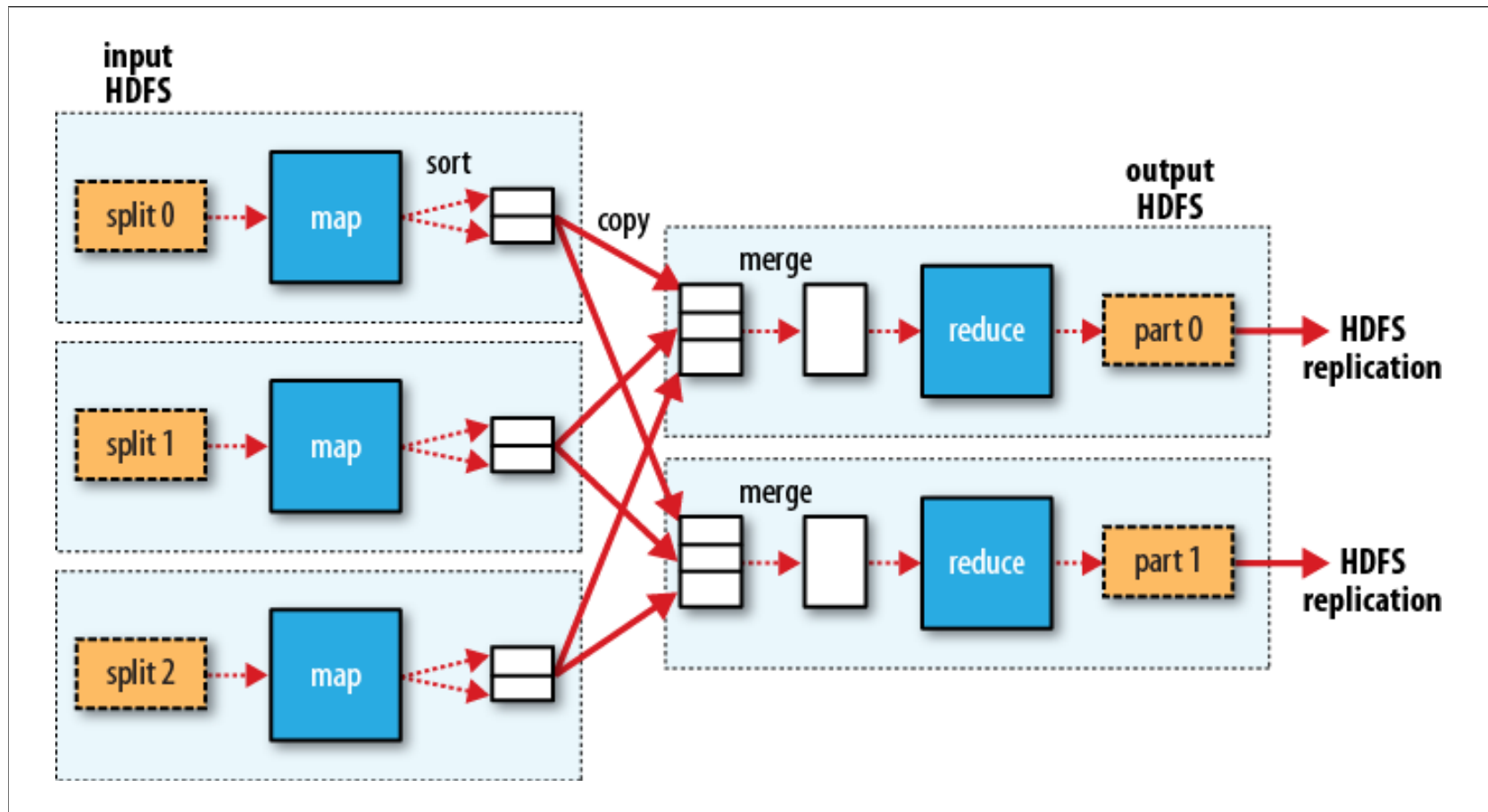
- Our goal is to aggregate user actions into sessions, so we can better understand
 - User behavior
 - The impact changes have on user behavior
- So what should a session look like?

User Session

- Data about the session as a whole
- List of events (pages viewed, actions taken)
 - Ordered in time
- In our logs, what data is session-wide
- What data is impression/action specific

MapReduce in Hadoop

Figure 2.4, Hadoop - The Definitive Guide



Assignment 5

- Define an Avro object for a user session
 - One user session for each unique userID
 - Session will include an array of events
 - Events ordered by timestamp
- Identify data associated with the session as a whole
- Identify data associated with individual events
- Include all the fields in the log entries
- Create enums for:
 - `body_style`, `cab_style`, `vehicle_condition`

Assignment 5

- Run WordCount on `dataSet5a.tsv`, `dataSet5b.tsv` – see what's in these files
 - Modify WordCount to output values for each field:
 - `fieldname:value`
 - Ignore these fields (they have lots of values):
 - `event_timestamp`, `price`, `mileage`, `user_id`, `vin`
- `event_type`
 - Break this into two fields in your schema:
 - `event_type` (enum), `event_subtype` (enum)

Assignment 5

Recommendations

- Get your app working with just a few fields populated
 - Session with no events, or just a count of events
 - Add events, but just a few fields first
 - Extend the schema
 - Populate the new field(s) in your schema
- Look at file `dataSet5Small.tsv` (on Canvas) to understand the data
- Write some unit tests as you go