CSE427s - 8 Lab8

50% (2/4)

✓ 1. **Flume** is used for...

- A batch data ingest from RDMS to HDFS
- B batch data ingest from a local filesystem to HDFS
- C ingesting streaming data into HDFS
- D stream processing

✗ 2. What are **short comings of Hadoop MapReduce** that are overcome by Spark?

- A serialization is limited to key-value paris
- B job workflows consisting of multiple jobs can only be DAGs (directed acyclic graphs)
- C job workflows consisting of multiple jobs read and write the output of each job to HDFS
- D data in HDFS is read-only (no random writes and updates)
- E Java only

⊘ 3. Name a Spark operation that is a **transformation.**

map()

⊘ 4. Name a Spark operation that is an **action.**

count()

✓ 5. What are properties of transformations and actions?

- A transformations return a variable/value, actions produce an RDD
- B actions return a variable/value, transformations produce an RDD
- C job execution is triggered by transformations
- D job execution is triggered by actions
- E the output of transformations is stored in HDFS
- F the output of actions is stored in HDFS

✗ 6. Why is the **Spark execution** so fast? (select all that apply)

- A Command Chaining
- B Lazy Execution
- C Pipelining

7. Post your spark code to **create the pair RDD** for postal codes in the form (postalcode, (lat,long)).

```
mydata.map(lambda x:(x.split(" ")[0],(x.split(" ")[1],x.split(" ")[2])))
```

8. Post your spark code to **create the pair RDD** for product ids and skus in the form (product_id, sku).