Introduction to Database Systems
CSE 414

Lecture 18: Spark

Data Model

Files!

A file = a bag of `(key, value)` pairs
Sounds familiar after HW5?

A MapReduce program:
- Input: a bag of `(inputkey, value)` pairs
- Output: a bag of `(outputkey, value)` pairs
  - outputkey is optional

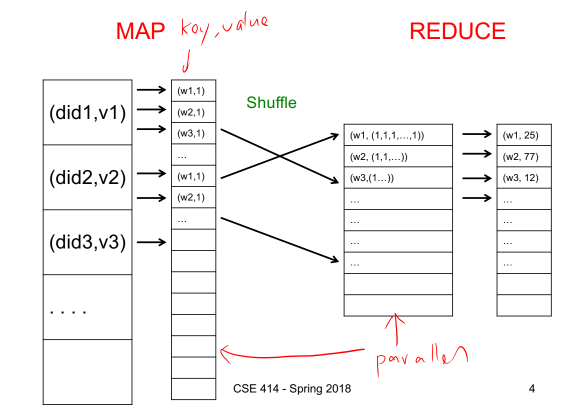
Example

- Counting the number of occurrences of each word in a large collection of documents
- Each Document
  - The key = document id (did)
  - The value = set of words (word)

```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
     emitIntermediate(w, "1");
```

key value

```
reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
    for each v in values:
      result += ParseInt(v);
  emit(AsString(result));
```



Fault Tolerance

- If one server fails once every year…
  ... then a job with 10,000 servers will fail in less than one hour
- MapReduce handles fault tolerance by writing intermediate files to disk:
  - Mappers write file to local disk
  - Reducers read the files (=reshuffling); if the server fails, the reduce task is restarted on another server

Group By $\gamma_{A,sum(B)}(R)$

map(Tuple t):
  EmitIntermediate(t.A, t.B);

(23, [10 (from t_1)])
(42, [6 (from t_4)])
(123, [21 (from t_2), 4 (from t_3)])

	A	B
t_1	23	10
t_2	123	21
t_3	123	4
t_4	42	6

reduce(String A, Iterator values):
  s = 0
  for each v in values:
    s = s + v
  Emit(A, s);

(23, 10), (42, 6), (123, 25)

R(A,B) $\bowtie_{B=C}$ S(C,D)

Partitioned Hash-Join

actual tuple

```
map(Tuple t):
   case t.relationName of
        'R': EmitIntermediate(t.B, t);
        'S': EmitIntermediate(t.C, t);
```

(5, [t_1, t_3, t_4])
(6, [t_2, t_5])

```
reduce(String k, Iterator values):
   R = empty;  S = empty;
   for each v in values:
        case v.relationName of:
          'R':   R.insert(v)
          'S':   S.insert(v);
   for v1 in R, for v2 in S
        Emit(v1,v2);
```

(t_1, t_4), (t_3, t_4)
(t_2, t_5)

R

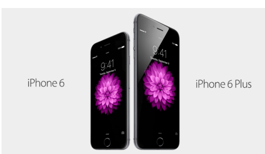
	A	B
t_1	10	5
t_2	20	6
t_3	20	5

S

	C	D
t_4	5	2
t_5	6	10

Spark

A Case Study of the MapReduce Programming Paradigm



Parallel Data Processing @ 2010

Issues with MapReduce

- Difficult to write more complex queries
- Need multiple MapReduce jobs: dramatically slows down because it writes all results to disk

Spark

- Open source system from UC Berkeley
- Distributed processing over HDFS
- Differences from MapReduce:
  - Multiple steps, including iterations
  - Stores intermediate results in main memory
  - Closer to relational algebra (familiar to you)
- Details: http://spark.apache.org/examples.html

Spark

- Spark supports interfaces in Java, Scala, and Python
  - Scala: extension of Java with functions/closures
- We will illustrate use the Spark Java interface in this class
- Spark also supports a SQL interface (SparkSQL), and compiles SQL to its native Java interface

Resilient Distributed Datasets

- RDD = Resilient Distributed Datasets
  - A distributed, immutable relation, together with its *lineage*
  - Lineage = expression that says how that relation was computed = a relational algebra plan
- Spark stores intermediate results as RDD
- If a server crashes, its RDD in main memory is lost. However, the driver (=master node) knows the lineage, and will simply recompute the lost partition of the RDD

Programming in Spark

- A Spark program consists of:
  - Transformations (map, reduce, join…). Lazy
  - Actions (count, reduce, save...). Eager
- Eager: operators are executed immediately
- Lazy: operators are not executed immediately
  - A *operator tree* is constructed in memory instead
  - Similar to a relational algebra tree

What are the benefits of lazy execution?

The RDD Interface

Collections in Spark

- `RDD<T>` = an RDD collection of type T
  - Partitioned, recoverable (through lineage), not nested
- `Seq<T>` = a sequence
  - Local to a server, may be nested

Example

Given a large log file `hdfs://logfile.log` retrieve all lines that:

- Start with "`ERROR`"
- Contain the string "`sqlite`"

```
s = SparkSession.builder()...getOrCreate();
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l -> l.startsWith("ERROR"));
sqlerrors = errors.filter(l -> l.contains("sqlite"));
sqlerrors.collect();
```

Example

Given a large log file `hdfs://logfile.log` retrieve all lines that:

- Start with "`ERROR`"
- Contain the string "`sqlite`"

`lines, errors, sqlerrors` have type `JavaRDD<String>`

```
s = SparkSession.builder()...getOrCreate();
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l -> l.startsWith("ERROR"));
sqlerrors = errors.filter(l -> l.contains("sqlite"));
sqlerrors.collect();
```

Example

Given a large log file `hdfs://logfile.log` retrieve all lines that:

- Start with "`ERROR`"
- Contain the string "`sqlite`"

`lines, errors, sqlerrors` have type `JavaRDD<String>`

```
s = SparkSession.build        e();
lines = s.read().textFi        gfile.log");
errors = lines.filter(l   l.startsWith("ERROR"));
sqlerrors = errors.filter(l        e"));
sqlerrors.collect();
```

Transformation: Not executed yet…

Action: triggers execution of entire program

Example

Recall: anonymous functions (lambda expressions) starting in Java 8

```
errors = lines.filter(l -> l.startsWith("ERROR"));
```

is the same as:

```
class FilterFn implements Function<Row, Boolean>{
  Boolean call (Row l)
  { return l.startsWith("ERROR"); }
}

errors = lines.filter(new FilterFn());
```

Example

Given a large log file `hdfs://logfile.log` retrieve all lines that:

- Start with "`ERROR`"
- Contain the string "`sqlite`"

```
s = SparkSession.builder()...getOrCreate();

sqlerrors = s.read().textFile("hdfs://logfile.log")
             .filter(l -> l.startsWith("ERROR"))
             .filter(l -> l.contains("sqlite"))
             .collect();
```

"Call chaining" style

MapReduce Again…

Steps in Spark resemble MapReduce:

- `col.filter(p)` applies in parallel the predicate p to all elements x of the partitioned collection, and returns collection with those x where `p(x) = true`
- `col.map(f)` applies in parallel the function f to all elements x of the partitioned collection, and returns a new partitioned collection

22

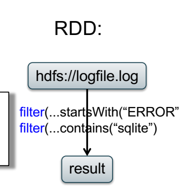
Persistence

```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect();
```

If any server fails before the end, then Spark must restart

Persistence

```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect();
```



If any server fails before the end, then Spark must restart

Persistence

RDD:


```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect();
```

If any server fails before the end, then Spark must restart

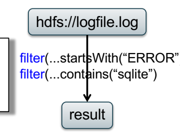
```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
errors.persist();
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect()
```

New RDD

Spark can recompute the result from errors

Persistence

RDD:

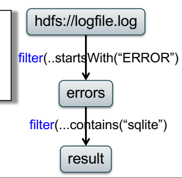



```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect();
```

If any server fails before the end, then Spark must restart

```
lines = s.read().textFile("hdfs://logfile.log");
errors = lines.filter(l->l.startsWith("ERROR"));
errors.persist();
sqlerrors = errors.filter(l->l.contains("sqlite"));
sqlerrors.collect()
```

New RDD



Spark can recompute the result from errors

R(A,B)
S(A,C)

SELECT count(*) FROM R, S
WHERE R.B > 200 and S.C < 100 and R.A = S.A

Example

```
R = s.read().textFile("R.csv").map(parseRecord).persist();
S = s.read().textFile("S.csv").map(parseRecord).persist();
```

Parses each line into an object

persisting on disk

R(A,B)
S(A,C)

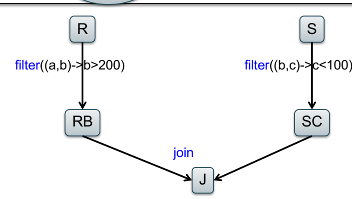
SELECT count(*) FROM R, S
WHERE R.B > 200 and S.C < 100 and R.A = S.A

Example

```
R = s.read().textFile("R.csv").map(parseRecord).persist();
S = s.read().textFile("S.csv").map(parseRecord).persist();
RB = R.filter(t -> t.b > 200).persist();
SC = S.filter(t -> t.c < 100).persist();
J = RB.join(SC).persist();
J.count();
```

transformations

action



Recap: Programming in Spark

- A Spark/Scala program consists of:
  - Transformations (map, reduce, join…). Lazy
  - Actions (count, reduce, save...). Eager
- RDD<T> = an RDD collection of type T
  - Partitioned, recoverable (through lineage), not nested
- Seq<T> = a sequence
  - Local to a server, may be nested

Transformations:	
`map(f : T -> U):`	`RDD<T> -> RDD<U>`
`flatMap(f: T -> Seq(U)):`	`RDD<T> -> RDD<U>`
`filter(f:T->Bool):`	`RDD<T> -> RDD<T>`
`groupByKey():`	`RDD<(K,V)> -> RDD<(K,Seq[V])>`
`reduceByKey(F:(V,V)-> V):`	`RDD<(K,V)> -> RDD<(K,V)>`
`union():`	`(RDD<T>,RDD<T>) -> RDD<T>`
`join():`	`(RDD<(K,V)>,RDD<(K,W)>) -> RDD<(K,(V,W))>`
`cogroup():`	`(RDD<(K,V)>,RDD<(K,W)>)-> RDD<(K,(Seq<V>,Seq<W>))>`
`crossProduct():`	`(RDD<T>,RDD<U>) -> RDD<(T,U)>`

Actions:	
`count():`	`RDD<T> -> Long`
`collect():`	`RDD<T> -> Seq<T>`
`reduce(f:(T,T)->T):`	`RDD<T> -> T`
`save(path:String):`	Outputs RDD to a storage system e.g., HDFS

Spark 2.0

The DataFrame and Dataset Interfaces

DataFrames

- Like RDD, also an immutable distributed collection of data
- Organized into *named columns* rather than individual objects
  - Just like a relation
  - Elements are untyped objects called `Row`'s
- Similar API as RDDs with additional methods
  - `people = spark.read().textFile(…);`
    `ageCol = people.col("age");`
    `ageCol.plus(10); // creates a new DataFrame`

Datasets

- Similar to DataFrames, except that elements must be typed objects
- E.g.: `Dataset<People>` rather than `Dataset<Row>`
- Can detect errors during compilation time
- DataFrames are aliased as `Dataset<Row>` (as of Spark 2.0)
- You will use both Datasets and RDD APIs in HW6

Datasets API: Sample Methods

- Functional API
  - `agg(Column expr, Column... exprs)`
    Aggregates on the entire Dataset without groups.
  - `groupBy(String col1, String... cols)`
    Groups the Dataset using the specified columns, so that we can run aggregation on them.
  - `join(Dataset<?> right)`
    Join with another DataFrame.
  - `orderBy(Column... sortExprs)`
    Returns a new Dataset sorted by the given expressions.
  - `select(Column... cols)`
    Selects a set of column based expressions.
- "SQL" API
  - `SparkSession.sql("select * from R");`
- Look familiar?

Conclusions

- Parallel databases
  - Predefined relational operators
  - Optimization
  - Transactions
- MapReduce
  - User-defined map and reduce functions
  - Must implement/optimize manually relational ops
  - No updates/transactions
- Spark
  - Predefined relational operators
  - Must optimize manually
  - No updates/transactions

35