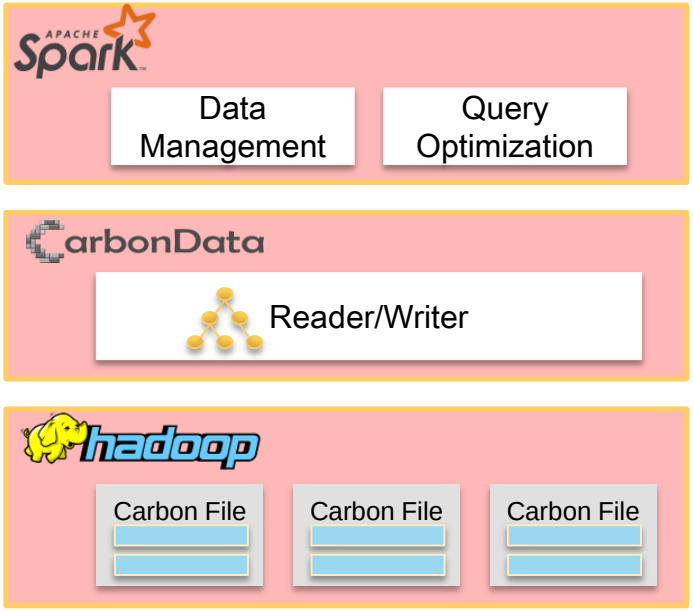
CarbonData: Spark Integration And Carbon Query Flow

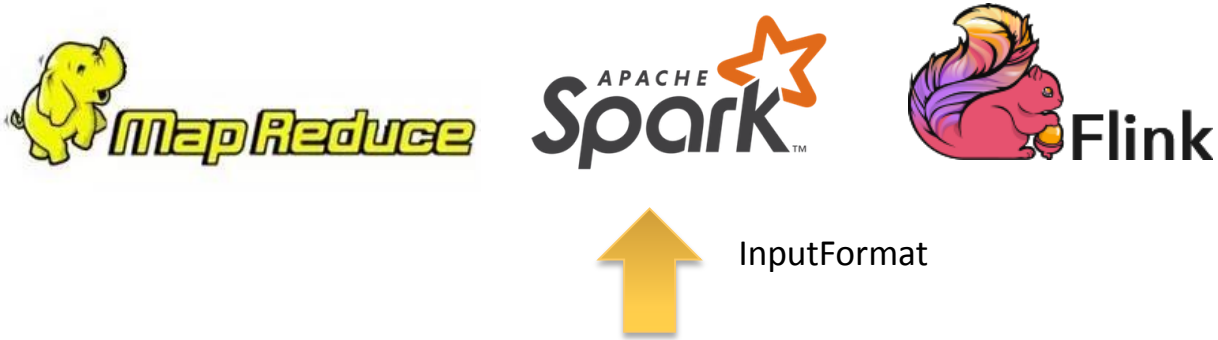
SparkSQL + CarbonData:

Carbon-Spark Integration

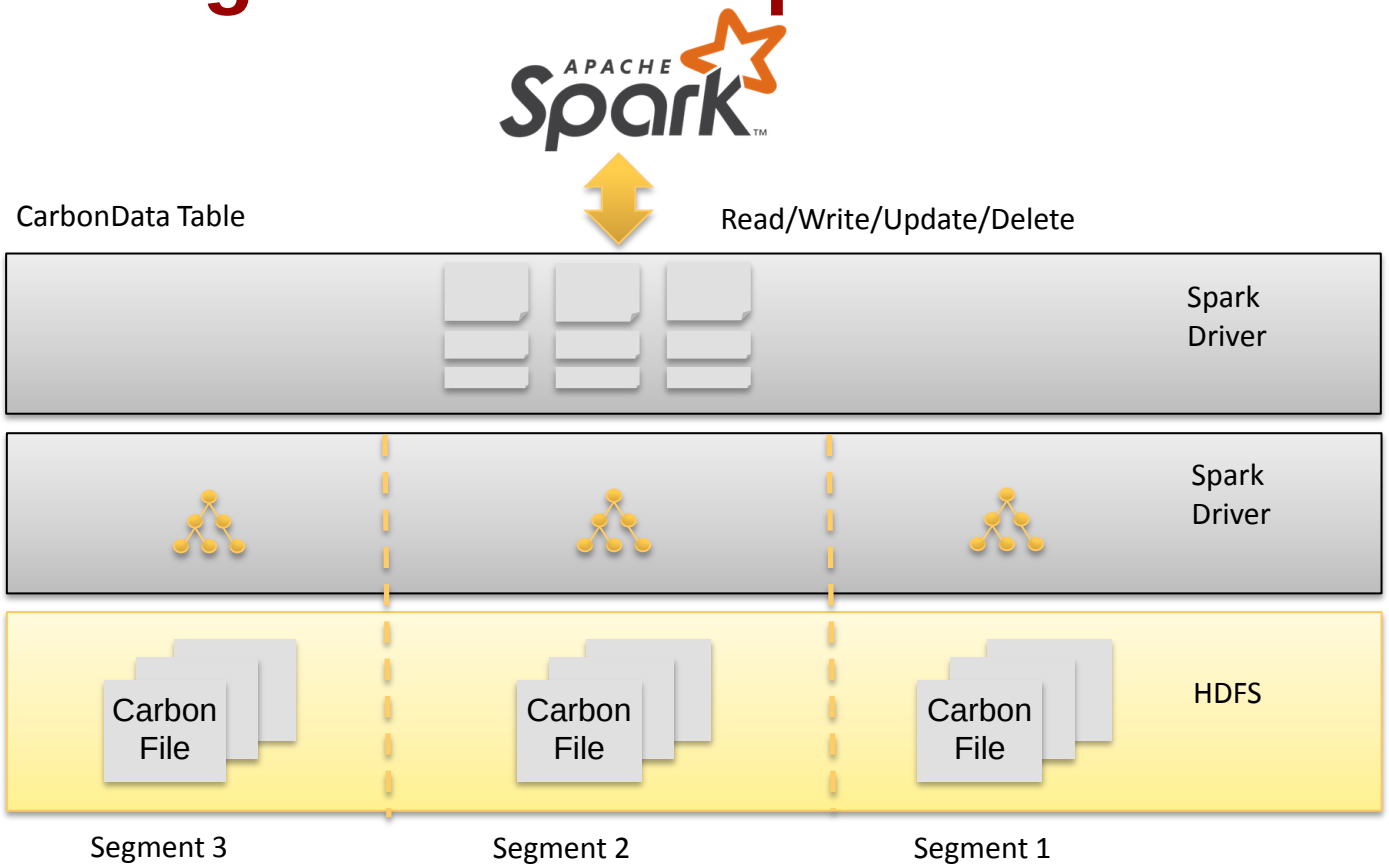
- **Built-in Spark integration**
  - Spark 1.5, 1.6, 2.1
- **Interface**
  - SQL
  - DataFrame API
- **Integration:**
  - File Format
  - Query Optimization & Data Management



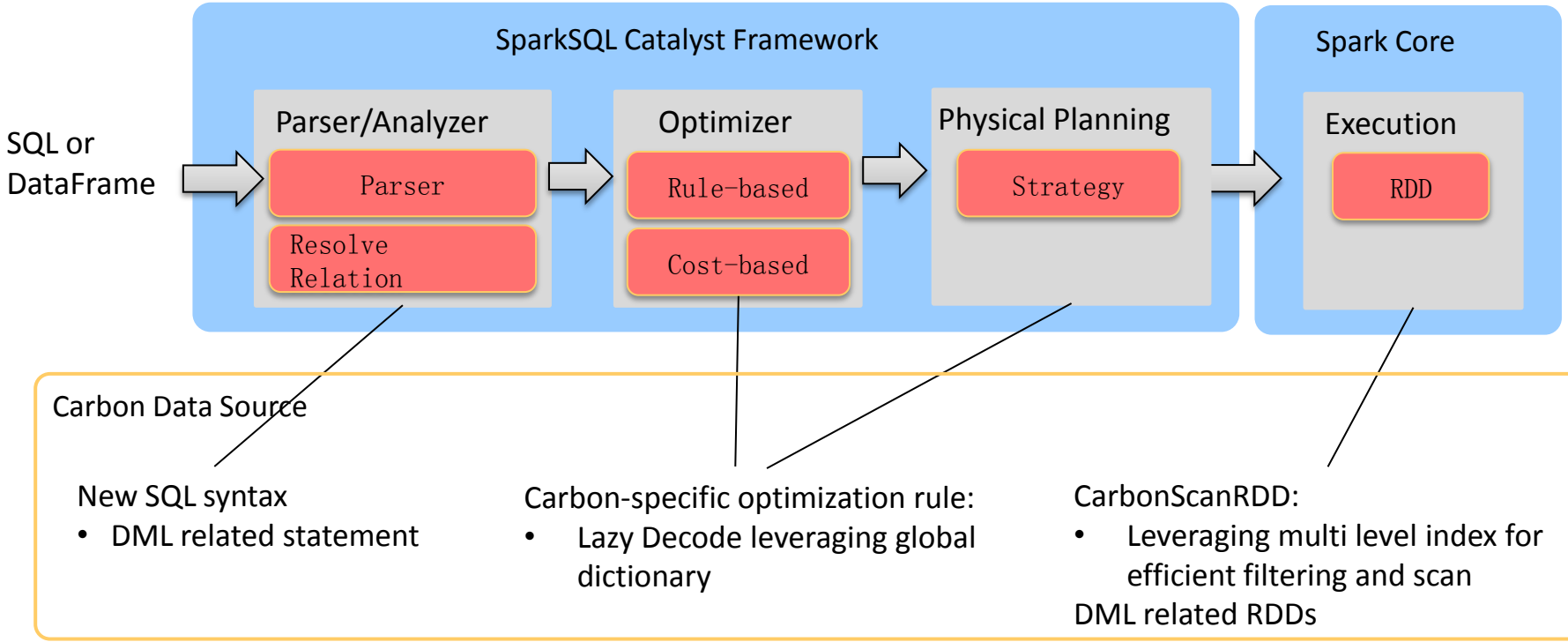
Integration with MR



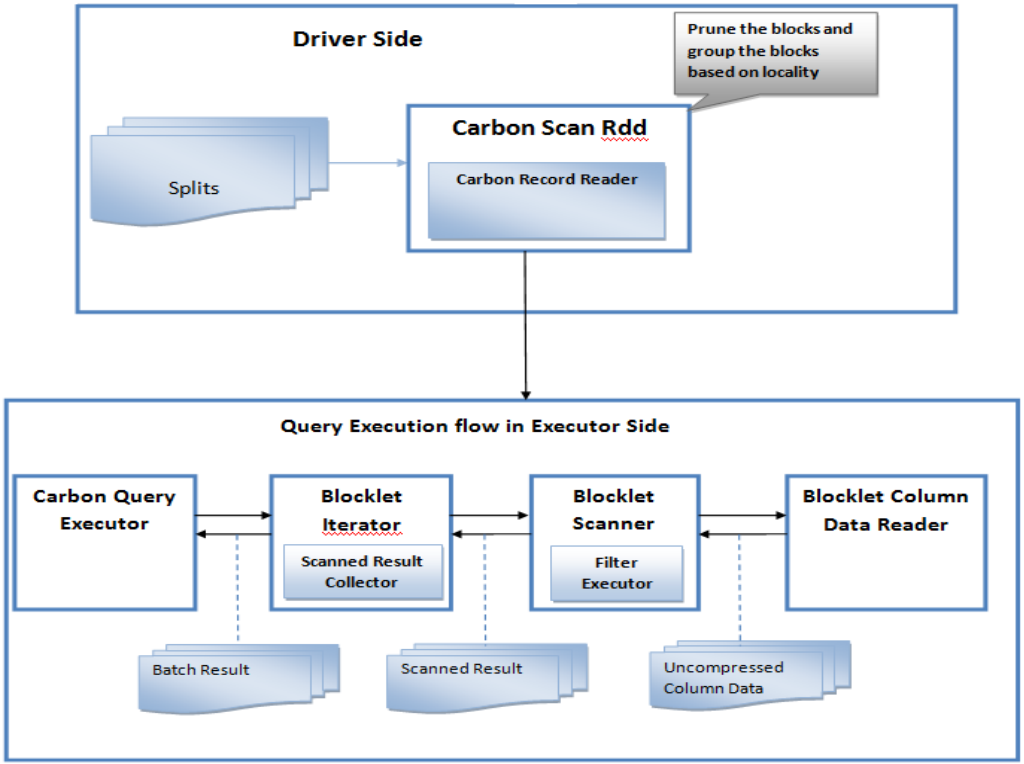
Deep Integration with Spark



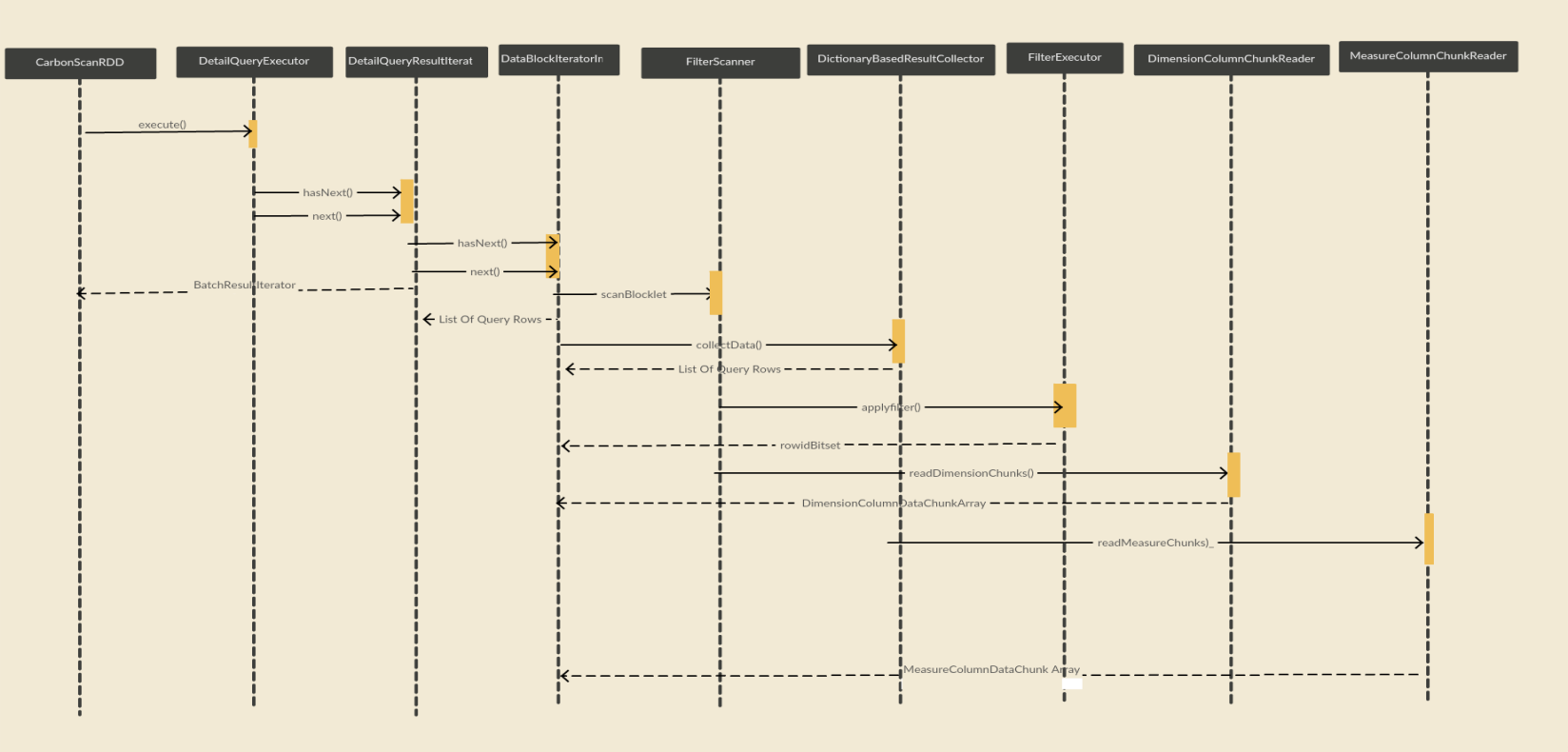
CarbonData as a SparkSQL Data Source



Query Flow Diagram

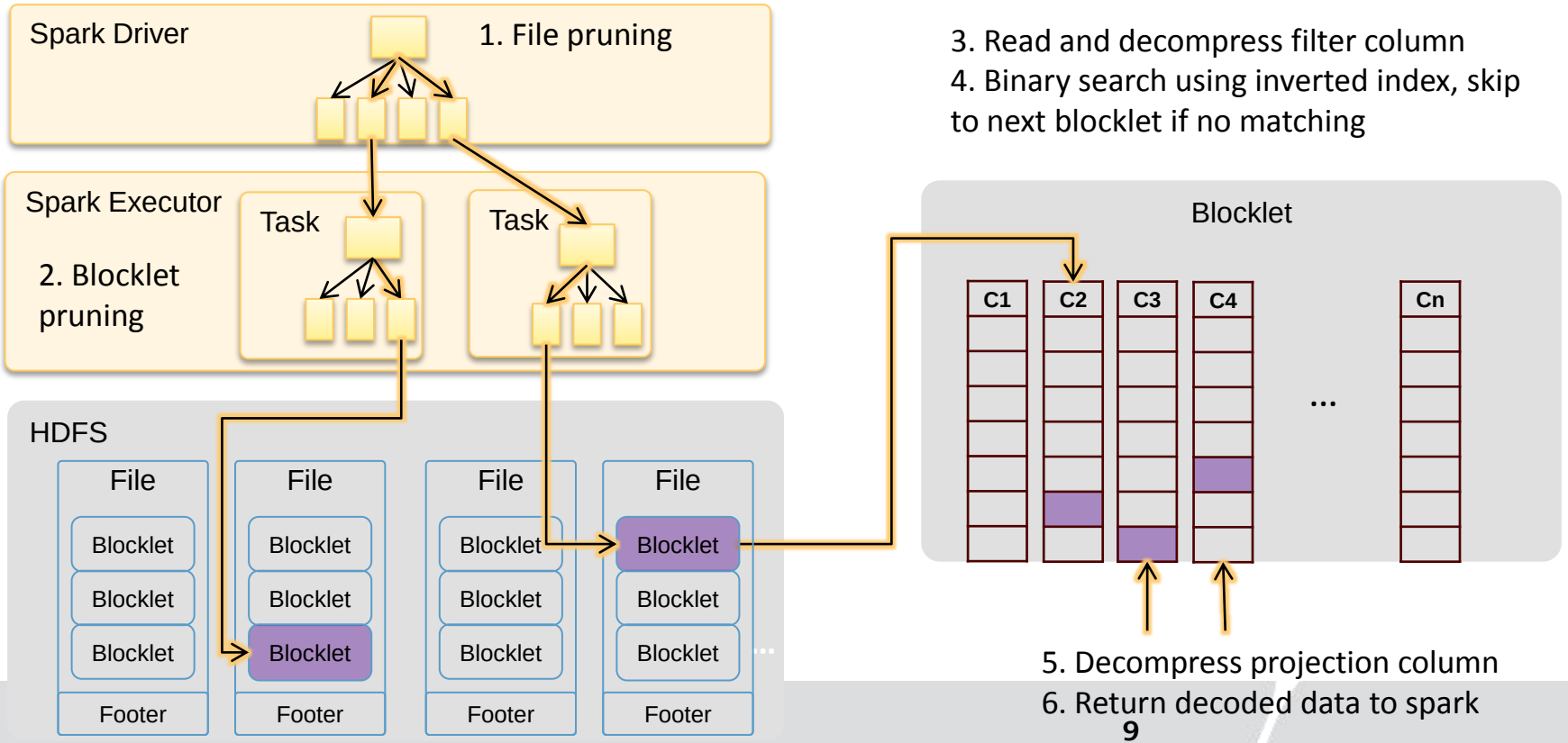


Query Sequence Diagram



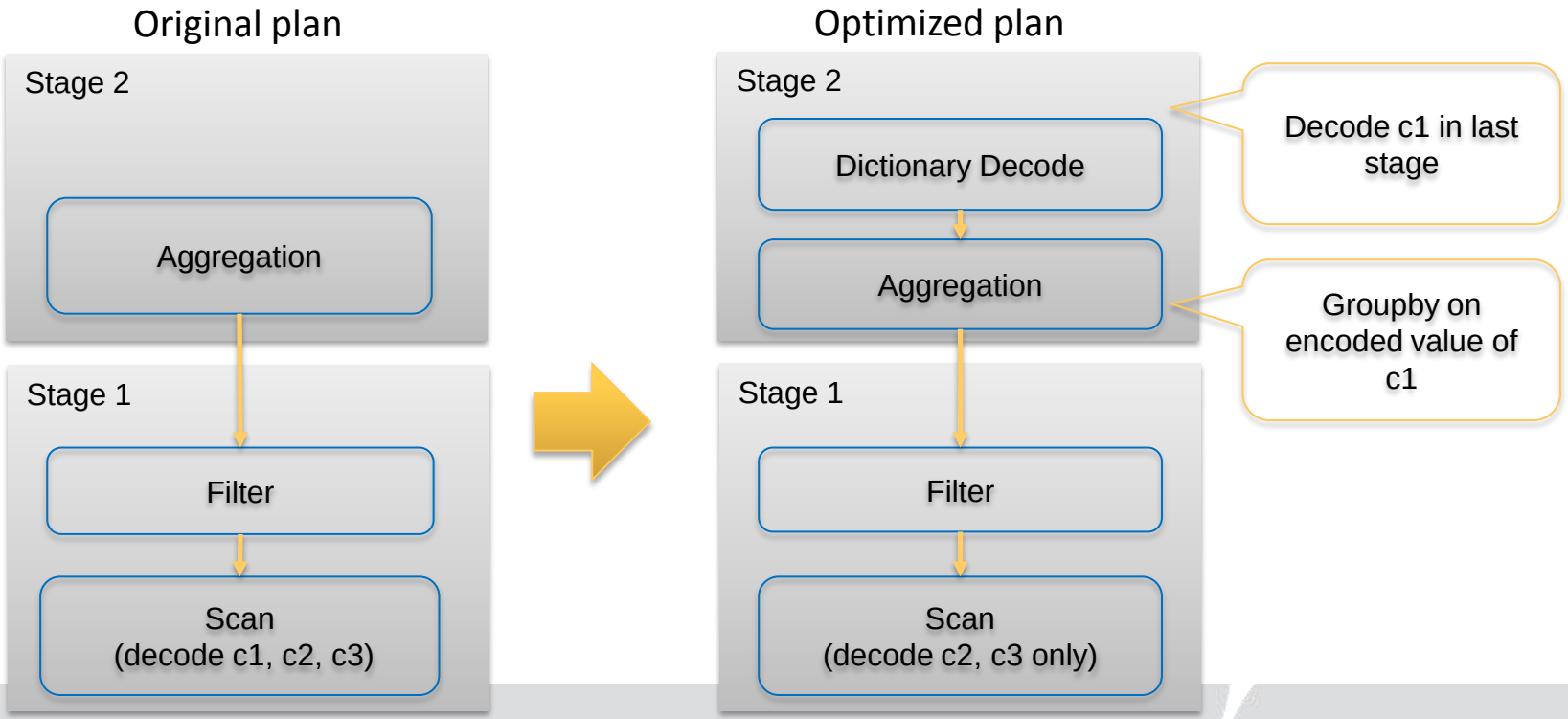
Efficient Filtering via Index

```
SELECT c3, c4 FROM t1 WHERE c2='boston'
```



Lazy Decoding by Leveraging Global Dictionary

```
SELECT c1, sum(c2) FROM t1 WHERE c3>10 group by c1
```



Query Package And Interface

Query RDD

•*org.apache.carbondata.spark.rdd.CarbonScanRdd*-Query execution rdd

Query Executor Interface

•*org.apache.carbondata.core.scan.executor.QueryExecutor*- Carbon query interface

•*org.apache.carbondata.core.scan.result.iterator.AbstractDetailQueryResultIterator*- Internal query interface, execute the query return the iterator over query result

Scanner Interface

•*org.apache.carbondata.core.scan.processor.AbstractDataBlockIterator*- Blocklet iterator to process the blocklet

Query Package And Interface

Scanner Interface

•*org.apache.carbondata.core.scan.scanner.BlockletScanner*- Interface for scanning the blocklet, there are two type of scanner non filter scanner and filter scanner.

•*org.apache.carbondata.core.scan.filter.executer.FilterExecuter* -Interface for executing the filter in executor side

•*org.apache.carbondata.core.scan.collector.ScannedResultCollector* - Prepares the query results from scanned result

Reader Interface

•*org.apache.carbondata.core.datastore.chunk.reader.DimensionColumnChunkReader*- Reader interface for reading and uncompressing the blocklet dimension column data.
Reader returns the store instance based on column type

Query Package And Interface

Reader Interface

•*org.apache.carbondata.core.datastore.chunk.reader.MeasureColumnChunkReader*- Reader interface for reading and uncompressing the blocklet measure column data.
Reader returns the store instance based on column data type

Query Data Store Interface

•*org.apache.carbondata.core.datastore.chunk.store.DimensionDataChunkStore*- Interface for storing the uncompressed dimension column data during query

•*org.apache.carbondata.core.datastore.chunk.store. MeasureDataChunkStore-* Interface for storing the uncompressed measure column data during query

Query Package And Interface

Dictionary Decoder

- *org.apache.spark.sql.CarbonDictionaryDecoder-* Decodes the dictionary values to actual data

Thank you