|  |
| --- |
| FICS |
| Hadoop |
| MapReduce |

|  |
| --- |
| SeokRae Kim  2017-7-31 |

목 차

[**I.** **MapReduce Application 개발** 1](#_Toc489906725)

[**1.** **항공 데이터 분석** 1](#_Toc489906726)

[**2.** **공통 클래스(AirlinePerformanceParser) 구현** 5](#_Toc489906727)

[**3.** **항공 출발 지연 데이터 분석** 5](#_Toc489906728)

[**4.** **항공 도착 지연 데이터 분석** 9](#_Toc489906729)

[**5.** **사용자 정의 옵션 사용** 12](#_Toc489906730)

[**6.** **카운터 사용** 15](#_Toc489906731)

[**7.** **다수의 파일 출력** 17](#_Toc489906732)

[**II.** **정렬** 20](#_Toc489906733)

[**1.** **보조 정렬** 20](#_Toc489906734)

[**2.** **부분 정렬** 22](#_Toc489906735)

1. **MapReduce Application 개발**
   1. **항공 데이터 분석**

* 데이터 다운로드를 위한 셸스크립트 작성
* CSV 파일 HDFS에 업로드
* 공통 클래스(AirlinePerformanceParser) 작성
* 항공 출발 지연 데이터 분석을 위한 매퍼(DepartureDelayCountMapper) 작성
* 매퍼에서 출력한 데이터를 처리하는 리듀서(DelayCountReducer) 작성
* 매퍼와 리듀서를 실행하는 드라이버(DepartureDelayCount) 작성
* 드라이버 실행

|  |
| --- |
| **#데이터 내려 받기** |
| **http://stat-computing.org/dataexpo/2009/**    **#1987년 ~ 2008년까지 총 12GB의 데이터로 하둡을 테스트하기로 한다.** |

|  |
| --- |
| **#경로 만들기** |
| **[root@wikibooks01 hadoop]# mkdir dataexpo** |

|  |
| --- |
| **#다운로드 받기 위한 셸 스크립트 작성** |
| **[root@wikibooks01 dataexpo]# vi downloads.sh**  **#!/bin/sh**  **for((i = 1987; i <= 2008; i++)) ; do**  **wget http://stat-computing.org/dataexpo/2009/$i.csv.bz2**  **bzip2 -d $i.csv.bz2**  **sed -e '1d' $i.csv > $i\_temp.csv**  **mv $i\_temp.csv $i.csv**  **done** |

|  |
| --- |
| **#셸스크립트 실행 (다운 받는데 시간이 오래 걸림)** |
| **[root@wikibooks01 dataexpo]# chmod 755 downloads.sh**  **[root@wikibooks01 dataexpo]# ./downloads.sh** |

|  |
| --- |
| **#다운로드 확인 및 용량 확인** |
| **[root@wikibooks01 dataexpo]# ls –al**  **[root@wikibooks01 dataexpo]# du -h** |

|  |
| --- |
| **#CSV 파일을 HDFS에 업로드 및 확인 – 오래 걸림** |
| **[root@wikibooks01 dataexpo]# hadoop dfs –mkdir airdata**  **[root@wikibooks01 dataexpo]# hadoop dfs –put /home/hadoop/dataexpo/\*.csv airdata**  **[root@wikibooks01 dataexpo]# hadoop dfs –ls** |

지금까지 항공 데이터 분석을 위한 데이터를 모두 다운로드 받았다.

이 데이터를 통해 앞으로 “항공 출발 지연 데이터 분석”과 “항공 도착 지연 데이터 분석”을 실행한다.

이제 데이터 분석을 위해 필요한 프로그램 개발을 설명한다.

* 1. **공통 클래스(AirlinePerformanceParser) 구현**

|  |
| --- |
| **#프로그램 구조** |
| **chapter05**  - 매퍼(Mapper)  - 리듀서(Reducer)  - 드라이버(Driver) **common**  - Parser |

|  |
| --- |
| **#항공 데이터를 파싱 할 클래스(AirlinePerformanceParser) 구현** |
| <https://github.com/SeokRae/hadoop_beginner/tree/SeokRae-branch/src/main/java/wikibooks/hadoop/common> |

* 1. **항공 출발 지연 데이터 분석**
     1. **매퍼(DepartureDelayCountMapper) 구현**

|  |
| --- |
| **#DepartureDelayCountMapper 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DepartureDelayCountMapper.java> |

* + 1. **리듀서(DelayCountReducer) 구현**

|  |
| --- |
| **#DelatCountReducer 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountReducer.java> |

* + 1. **드라이버(DepartureDelayCount) 클래스 구현**

|  |
| --- |
| **#DepartureDelayCount 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DepartureDelayCount.java> |

* + 1. **드라이버 클래스 실행**

|  |
| --- |
| **#맵리듀스 잡 실행 (10분정도 소모)** |
| 실행 전 Maven build로 jar파일 생성을 잊지 말 것.  **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.DepartureDelayCount airdata dep\_delay\_count**    매퍼에서 데이터를 받아 리듀서에서 최종적으로 생성한 데이터를 보여주는 핵심 로그  **17/07/31 21:31:22 INFO mapred.JobClient: Map input records=123534969**  **17/07/31 21:31:22 INFO mapred.JobClient: Map output records=50018329**  **17/07/31 21:31:22 INFO mapred.JobClient: Reduce input records=50018329**  **17/07/31 21:31:22 INFO mapred.JobClient: Reduce input groups=255** |

|  |
| --- |
| **#HDFS에 저장된 출력 데이터확인** |
| **[root@wikibooks01 hadoop]# hadoop fs –ls**  **[root@wikibooks01 hadoop]# hadoop fs -cat dep-delay-count/part-r-00000 | head -10**  **[root@wikibooks01 hadoop]# hadoop fs -cat dep-delay-count/part-r-00000 | tail -10** |

|  |
| --- |
| **#엑셀로 시각화하기 위해 CSV 파일로 만들기** |
| **[root@wikibooks01 hadoop]# hadoop fs -get dep\_delay\_count/part-r-00000 dep\_delay\_count.csv**  dep\_delay\_count.csv 파일이 만들어졌는지 확인한다. |

|  |
| --- |
| **#엑셀 피벗차트를 이용해서 데이터 확인** |
| **CentOS에 있는 LibreOffice Calc를 이용해서 데이터를 읽은 후,**  **엑셀프로그램으로 피벗차트를 만들어 보았다.** |

* 1. **항공 도착 지연 데이터 분석**
     1. **매퍼(ArrivalDelayCountMapper) 구현**

|  |
| --- |
| **# ArrivalDelayCountMapper 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/ArrivalDelayCount.java> |

* + 1. **리듀서(DelayCountReducer) 구현**

|  |
| --- |
| **# DelayCountReducer 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountReducer.java> |

* + 1. **드라이버 클래스 구현**

|  |
| --- |
| **# ArrivalDelayCount 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/ArrivalDelayCount.java> |

* + 1. **드라이버 클래스 실행**

|  |
| --- |
| **#맵 리듀스 실행 (10분정도 소요)** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.ArrivalDelayCount airdata arr\_delay\_count**    **17/08/01 21:29:33 INFO mapred.JobClient: Map input records=123534969**  **17/08/01 21:29:33 INFO mapred.JobClient: Map output records=57839171**  **17/08/01 21:29:33 INFO mapred.JobClient: Reduce input records=57839171**  **17/08/01 21:29:33 INFO mapred.JobClient: Reduce output records=255** |

|  |
| --- |
| **#head 와 tail로 데이터 확인** |
| **[root@wikibooks01 hadoop]# hadoop fs -cat arr\_delay\_count/part-r-00000 | head -10**  **[root@wikibooks01 hadoop]# hadoop fs -cat arr\_delay\_count/part-r-00000 | tail -10** |

* + 1. **항공 도착 지연 데이터 차트 만들기**

|  |
| --- |
| **# CSV 파일 만들기** |
| **[root@wikibooks01 hadoop]# hadoop fs -get arr\_delay\_count/part-r-00000 arr \_delay\_count.csv** |

|  |
| --- |
| **#엑셀 피벗차트를 이용해서 데이터 확인** |
| **CentOS에 있는 LibreOffice Calc를 이용해서 데이터를 읽은 후,**  **엑셀프로그램으로 피벗차트를 만들어 보았다.** |

* 1. **사용자 정의 옵션 사용**
* 사용자가 정의한 파라미터를 통해 **도착** 혹은 **출발** 지연 데이터를 분석
  + 1. **매퍼(DelayCountMapper) 구현**

|  |
| --- |
| **# DelayCountMapper 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountMapper.java> |

* + 1. **리듀서 구현**

|  |
| --- |
| **#** **DelayCountReducer 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountReducer.java> |

* + 1. **드라이버 클래스 구현**

|  |
| --- |
| **# DelayCount 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCount.java> |

* + 1. **드라이버 클래스 실행**
       1. **workType = departure**

|  |
| --- |
| **#departure 파라미터를 이용해 출발 데이터를 분석** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.DelayCount -D workType=departure airdata departure\_delay\_count** |

|  |
| --- |
| **# 결과물 확인** |
| **3. 항공 출발 지연 데이터 분석과 같은 결과 인지를 확인**  **17/08/02 15:54:50 INFO mapred.JobClient: Map input records=123534969**  **17/08/02 15:54:50 INFO mapred.JobClient: Map output records=50018329**  **17/08/02 15:54:50 INFO mapred.JobClient: Reduce input records=50018329**  **17/08/02 15:54:50 INFO mapred.JobClient: Reduce output records=255** |

|  |
| --- |
| **#HDFS에 저장된 데이터 확인** |
| **[root@wikibooks01 hadoop]# hadoop fs -cat departure\_delay\_count/part-r-00000 | head -10**  **[root@wikibooks01 hadoop]# hadoop fs -cat departure\_delay\_count/part-r-00000 | tail -10** |

* + - 1. **workType = arrival**

|  |
| --- |
| **#arrival 파라미터를 이용해 도착 지연 데이터 분석** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.DelayCount -D workType=arrival airdata arrival\_delay\_count**    **4. 항공 도착 지연 데이터 분석의 결과와 동일한지 확인**  **17/08/02 16:17:59 INFO mapred.JobClient: Map input records=123534969**  **17/08/02 16:17:59 INFO mapred.JobClient: Map output records=57839171**  **17/08/02 16:17:59 INFO mapred.JobClient: Reduce input records=57839171**  **17/08/02 16:17:59 INFO mapred.JobClient: Reduce output records=255** |

* 1. **카운터 사용**
* **Job 실행 결과 메시지와 웹 인터페이스에서 간편하게 데이터를 확인하기 위한 작업**
  + 1. **사용자 정의 카운터 구현**

|  |
| --- |
| **# DelayCounters 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCounters.java> |

* + 1. **매퍼 구현**

|  |
| --- |
| **# DelayCountMapperWithCounter 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountMapperWithCounter.java> |

* + 1. **드라이버 클래스 구현**

|  |
| --- |
| **# DelayCountWithCounter 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountWithCounter.java> |

* + 1. **드라이버 클래스 실행**

|  |
| --- |
| **# 맵리듀스 실행하기** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.DelayCountWithCounter -D workType=departure airdata departure\_delay\_count\_counter**    **도착 지연을 확인할 수 없는 경우 2,302, 136건**  **스케줄 대로 도착한 경우 26,416,798건**  **스케줄 보다 일찍 도착한 경우 44,797,706건** |

|  |
| --- |
| **#맵리듀스 잡 관리자용 웹 화면** |
| **DelayCounters 확인** |

* 1. **다수의 파일 출력**
     1. **MultipleOutputs 이해**

MultipleOutputs는 여러 개의 출력 데이터를 쉽게 생성하도록 돕는 기능을 제공합니다. MultipleOutputs는 여러 개의 OutputCollectors를 만들고 각 OutputCollectors에 대한 출력 경로, 출력 포맷, 키와 값 유형을 설정합니다.

* + 1. **매퍼 구현**

|  |
| --- |
| **#** **DelayCountMapperWithMultipleOutputs 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountMapperWithMultipleOutputs.java> |

* + 1. **리듀서 구현**

|  |
| --- |
| **# DelayCountReducerWithMultipleOutputs 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountReducerWithMultipleOutputs.java> |

* + 1. **드라이버 클래스 구현**

|  |
| --- |
| **# DelayCountWithMultipleOutputs 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter05/DelayCountWithMultipleOutputs.java> |

* + 1. **드라이버 클래스 실행**

|  |
| --- |
| **#전체 카운터 정보를 출력** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-airportdata-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter05.DelayCountWithMultipleOutputs airdata delay\_count\_mos** |

|  |
| --- |
| **#출력 데이터 확인** |
| **[root@wikibooks01 hadoop]# hadoop fs -ls delay\_count\_mos**  **arrival-r-00000, departure-r-00000 파일 확인**    **\*\* part-r-00000 파일의 크기가 0인 이유는 reduce 메서드에서 context 객체에 write를 하지 않았기 때문이다.** |

|  |
| --- |
| **#출발 지연 데이터 확인** |
| **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_mos/departure-r-00000 | head -10**  **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_mos/departure-r-00000 | tail-10** |

|  |
| --- |
| **#도착 지연 데이터 확인** |
| **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_mos/arrival-r-00000 | head -10**  **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_mos/arrival-r-00000 | tail-10** |

1. **정렬**

* 항공 데이터 분석 프로그램에 보조 정렬을 적용
* 부분 정렬과 전체 정렬을 이용해 단순하게 입력 데이터만 정렬하는 맵리듀스 프로그램 개발
  1. **보조 정렬**
* 키의 값들을 그룹핑하고, 그룹핑된 레코드에 순서를 부여하는 방식

|  |
| --- |
| **#보조 정렬의 알고리즘** |
| 1. 기존 키의 값들을 조합한 복합 키(Composite Key)를 정의, 이때 키의 값 중에서 어떤 키를 그룹핑 키로 사용할지 결정 2. 복합키의 레코드를 정렬하기 위한 비교기(Comparator)를 정의 3. 그룹핑 키를 파티셔닝할 파티셔너(Partitioner)를 정의 4. 그룹핑 키를 정렬하기 위한 비교기(Comparator)를 정의 |

* + 1. **복합키 구현**

|  |
| --- |
| **# DataKey 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/DateKey.java> |

* + 1. **복합키 비교기 구현**

|  |
| --- |
| **# DataKeyComparator 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/DateKeyComparator.java> |

* + 1. **그룹키 파티셔너 구현**

|  |
| --- |
| **# GroupKeyPartitioner 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/GroupKeyPartitioner.java> |

* + 1. **그룹키 비교기 구현**

|  |
| --- |
| **# GroupKeyComparator 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/GroupKeyComparator.java> |

* + 1. **매퍼 구현**

|  |
| --- |
| **# DelayCountMapperWithDateKey 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/DelayCountMapperWithDateKey.java> |

* + 1. **리듀서 구현**

|  |
| --- |
| **# DelayCountReducerWithDateKey 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/DelayCountReducerWithDateKey.java> |

* + 1. **드라이버 구현**

|  |
| --- |
| **# DelayCountWithDateKey 작성** |
| <https://github.com/SeokRae/hadoop_beginner/blob/SeokRae-branch/src/main/java/wikibooks/hadoop/chapter06/DelayCountWithDateKey.java> |

* + 1. **드라이버 실행**

|  |
| --- |
| **# 드라이버 클래스 실행** |
| **[root@wikibooks01 hadoop]# hadoop jar hadoop-beginner-sort-0.0.1-SNAPSHOT.jar wikibooks.hadoop.chapter06.DelayCountWithDateKey airdata delay\_count\_sort** |

|  |
| --- |
| **# 출력 데이터 조회** |
| **[root@wikibooks01 hadoop]# hadoop fs –ls delay\_count\_sort** |

|  |
| --- |
| **# head, tail 데이터 조회** |
| **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_sort/arrival-r-00000 | head -15**  **[root@wikibooks01 hadoop]# hadoop fs –cat delay\_count\_sort/arrival-r-00000 | tail -15** |

* 1. **부분 정렬**