빅데이터 이해 중간시험 풀이 (50 점 만점) 2018.4.26

1. 다음을 빈 칸을 기술하라. (2 점 x 15 개 = 30 점)

(1) 빅데이터의 3 가지 주요 특성 3V 는 규모(volume), 다양성(variety), (속도(velocity)) 이다.

(2) 맵리듀스(MapReduce) 알고리즘은 맵(Map), (셔플(Shuffle)), 리듀스(Reduce) 3 단계로 구성된다.

(3) 하둡 구성의 핵심적인 2 요소는 맵리듀스, (HDFS(Hadoop Distributed File System)) 이다.

(4) (YARN(Yet Another Resource Negotiator))은 하둡 2.0 부터 적용되어 클러스터의 각 응용에 자원을 할당하고 모니터링하는 클러스터 자원 관리자이다.

(5) 리눅스에서 파일의 권한을 변경하는 명령은 (chmod) 이다.

(6) 우분투에서는 패키지 관리의 설치 및 관리를 위해 (APT (Advanced Package Tool)) 소프트웨어 관리 도구를 사용한다.

(7) 하둡 맵리듀스의 (InputFormat) 클래스는 작업의 입력 데이터를 검증하고 맵 처리를 위해 파일을 분리한다.

(8) 하둡에서 마스터는 (네임 노드 (name node)), 슬레이브는 데이터 노드 (data node) 라 한다.

(9) 맵리듀스에서 마스터는 JobTracker, 슬레이브는 (TaskTracker) 라 한다.

(10) 맵리듀스의 슬레이브가 마스터에게 주기적으로 (heartbeats)을 전송하여 살아 있음을 알린다.

(11) 맵리듀스 프로그램을 위해서는 (드라이버(driver)), 매퍼(mapper), 리듀서(reducer) 클래스의 공통된 부분의 템플릿을 가지고 시작한다.

(12) 매퍼 클래스의 입력이 되는 디폴트 레코드 리더의 키는 데이터 파일의 (바이트 오프셋) 이다.

(13) 하둡이나 스파크의 실행 프로세스 확인 명령은 (jps) 이다.

(14) 함수형 프로그래밍 언어에서 (순수 함수(pure function))는 부작용(side-effect)이 없는 함수이다.

(15) 다른 함수를 매개 변수로 받아들이거나 반환값으로 함수를 사용하는 함수를 (고차 함수(high-order function)) 라고 한다.

2. 다음 스칼라 리스트의 실행 결과(리스트 또는 정수 값)를 기술하라. (5 개 X 2 점 = 10 점)

val nums = List(1,2,3,4,5,6,7)

(1) val n = nums.filter((i:Int) =>i % 3 == 0)

(2) n.map(_+1)

```
scala> val nums = List(1,2,3,4,5,6,7)
nums: List[Int] = List(1, 2, 3, 4, 5, 6, 7)

scala> val n = nums.filter((i:Int) =>i % 3 == 0)
n: List[Int] = List(3, 6)

scala> n.map(_+1)
res0: List[Int] = List(4, 7)
```

val colors = List("red", "green", "blue")

(3) val s = colors.map((c:String) => c.size)

(4) s.reduce((a:Int, b:Int) => a+b)

(5) colors.map(_.toList)

```
scala> val colors = List("red", "green", "blue")
colors: List[String] = List(red, green, blue)

scala> val s = colors.map( (c:String) => c.size)
s: List[Int] = List(3, 5, 4)

scala> s.reduce( (a:Int, b:Int) => a+b)
res1: Int = 12

scala> colors.map(_.toList)
res2: List[List[Char]] = List(List(r, e, d), List(g, r, e, e, n), List(b, l, u, e))
```

3. 아파치 스파크(Apache Spark)에서 다음을 기술하라. (10 점)

(1) 스파크의 자료 구조 RDD, 데이터세트 (Dataset), 데이터프레임(DataFrame)

RDD (Resilient Distributed Dataset)는 비구조화된 타입이 정의되지 않은 데이터로 프로그래머가 저수준 변환 및 액션 사용하며 최적화를 수행한다. 데이터세트는 반 기존의 최적화가 가능한 타입이 정의된 반-구조화/구조화된 데이터로 Spark SQL 의 프로그래밍 추상화 추가되었다. 데이터프레임은 이름을 갖는 열(named column)들로 구성된 반-구조화/구조화된 데이터로 스파크 SQL 최적화 실행 엔진을 사용하여 높은 최적화 수행한다.

(2) 스파크 연산의 2 가지 유형과 스파크 연산의 특징

스파크 연산은 크게 변환(transformation)가 액션(action)의 두 가지 연산을 수행한다. 변환 연산은 연산의 결과로 RDD 를 리턴한다. 액션 연산은 수행된 결과 값을 리턴하는 연산이다.

스파크는 데이터를 적재하는 즉시 RDD 를 생성하지 않고 효율적으로 사용자의 계산을 처리하는 방법을 찾기 위해 RDD 의 계산을 가능한 한 지연하고, 액션이 호출될 때 스파크는 실행 계획을 위해 만든 전체 그래프를 살펴보고 실행한다.