교육생용

DCX기반 빅데이터 분석 서비스 개발자과정

훈련생	(서명)	점수	

NCS 능력단위	2001020905_19v2 빅데이터 처리시스템 개발
유 형	이론형 (수기 작성)
제시조건	Question 1~25번 이론형 문제
평가영역 (NCS)	- 2001020905_19v2.1 빅데이터 처리시스템 설계하기 - 2001020905_19v2.2 빅데이터 처리시스템 구성하기 - 2001020905_19v2.3 분산 처리수행모듈 개발하기 - 2001020905_19v2.4 실시간 수행모듈 개발하기 - 2001020905_19v2.5 이벤트처리 수행모듈 개발하기
출제범위	분산처리시스템 하둡구축
시험일자	2023년 09월 22일 (금요일) 16:50-17:40(50분)

출제자	신인호 (서명)
검수자	차준섭, 강성관

Smart Media
스마트미디어인재개발원

Question 1	빅데이터 처리 시스템 설계하기

문제 1-1 (NCS 1.1)	기존 기술로 처리하지 못하는 대량의 데이터를 분산 스토리지에 저장하고, 저장된 대량의 데이터를 빠르게 조회 및 분석할 수 있도록 분산 환경에서 병렬로 실행하여 빠르고 안정적으로 처리할 수 있는 시스템이다. 위 에서 서술한 시스템은 무엇을 설명한 것인가?

① 빅데이터 수집시스템　② 빅데이터 저장시스템
③ 빅데이터 처리시스템　④ 빅데이터 분석시스템

답 :

문제 1-2 (NCS 1.1)	빅데이터는 과거와는 달리 현재는 모바일 기기 및 다양한 IT 기기를 통하여 데이터가 생성되고 다양해지고 있다. 빅데이터의 처리기술의 필요성과 거리가 먼 것은 어느 것인가?

①데이터 볼륨의 증가
②데이터 발생속도의 증가
③데이터포맷 다양성의 증가
④기존시스템과 연계성

답 :

문제 1-3 (NCS 1.3)	빅데이터 처리 유형은 처리할 데이터 유형에 따라서 정적 (㉠) 와 (㉡)로 분류할 수 있고, 실시간 처리는 대량의 데이터 처리·분석을 위한 실시간 처리 와 (㉢) 기반의 (㉣)으로 분류할 수 있다. 특히 실시간 데이터를 처리하면서, 동시에 이벤트 처리와 연동하여 정해진 룰에 맞는 이벤트를 탐지하도록 구성할 수 있다. 보기에서 고르시오.

보기
a. 배치처리　b. 실시간처리　c. 룰(Rule)　d. 이벤트 처리 탐지방식

답 :
㉠-( )　㉡-( )　㉢-( )　㉣-( )

Question 2	빅데이터 처리 시스템 구성하기

문제 2-1 (NCS 2.1)	( )은 빅데이터를 처리할 수 있는 컴퓨터 클러스터에서 분산처리 애플리케이션을 지원하는 오픈 소프트웨어 프레임워크로써, 기본적으로 2개의 핵심요소를 갖고 있다. 분산 파일시스템과 맵 리듀스(MapReduce)로 구성되어 있다. ( )는 무엇을 설명하고 있는가?

①EcoSytem　②Hadoop　③YARN　④Zookeper

답 :

문제 2-2 (NCS 2.2)	MapReduce는 복잡도가 높은 프로그래밍 기법이 필요했고, 이는 업무 분석가 및 관리자들에게 빅데이터에 접근하는 것을 어렵게 만들었습니다. 이를 해결하기 위해 페이스북에서 SQL과 매우 유사한 방식으로 하둡 데이터에 접근성을 쉽게하도록 개발 되었습니다. 하둡 에코시스템 중에서 데이터를 모델링하고 프로세싱하는 경우 가장 많이 사용하는 데이터 웨어하우징용 솔루션으로 활용되는 에코스시템은 무엇인가?

①주키퍼(Zookeeper)　②HBASE　③하이브(HIVE)　④Flume(플럼)

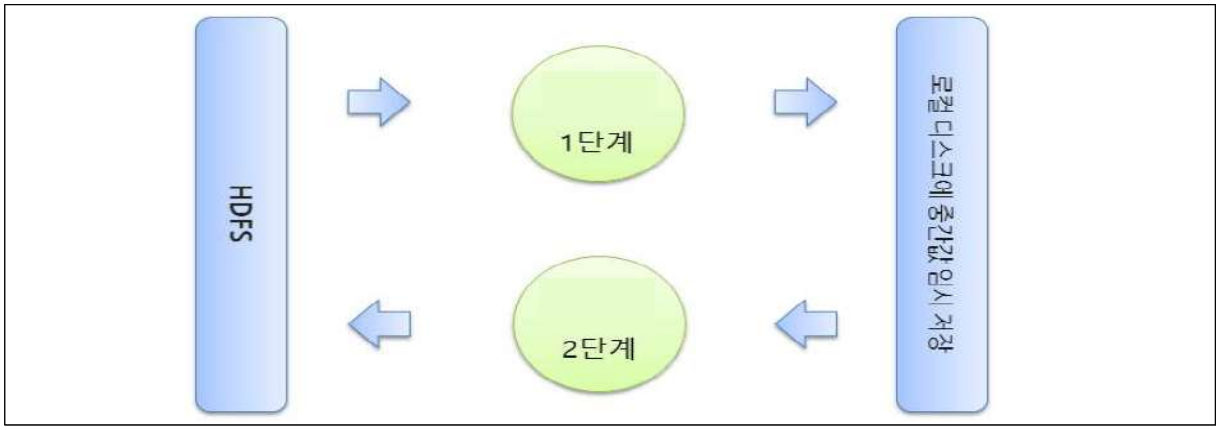
답 :

문제 2-3 (NCS 2.3)	실시간 대용량 데이터를 처리하는 관점에서 실시간으로 유입되는 데이터에서 패턴을 찾고, 탐지할 수 있도록 다양한 유형의 이벤트에서 특정 패턴을 발견할 수 있는 규칙(Rule)을 정의하고, 이를 기반으로 실시간 유입되는 데이터에서 가치를 찾거나, 위험 요인을 사전에 감지하여 대응할 수 있도록 시스템을 설계한다. 빅데이터 처리 모델 중 어느 것을 설명하는 글인가?

①분산처리 시스템　②배치처리 시스템
③실시간 처리 시스템　④이벤트 처리 시스템

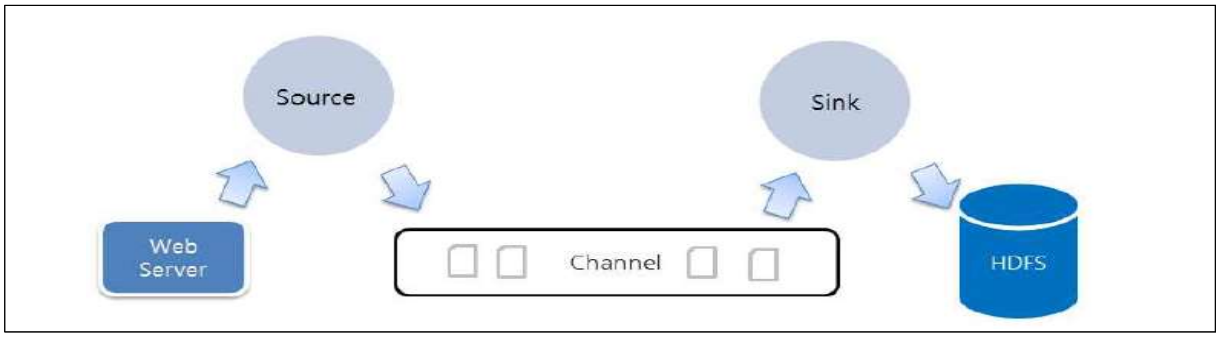
답 :

문제 2-4 (NCS 2.3)	아래 그림은 빅 데이터를 분산 처리하는 과정을 설명하고 있다. 1단계와 2단계에 알맞은 용어를 기술하시오



답 : 1단계 :
2단계 :

문제 2-5 (NCS 2.1)	아래 그림은 하둡 에코 시스템중에서 데이터를 수집하는 대표적인 오픈소스 소프웨어이다. 하둡의 에코 시스템 중 아래 그림은 무엇을 설명하고 있는가?



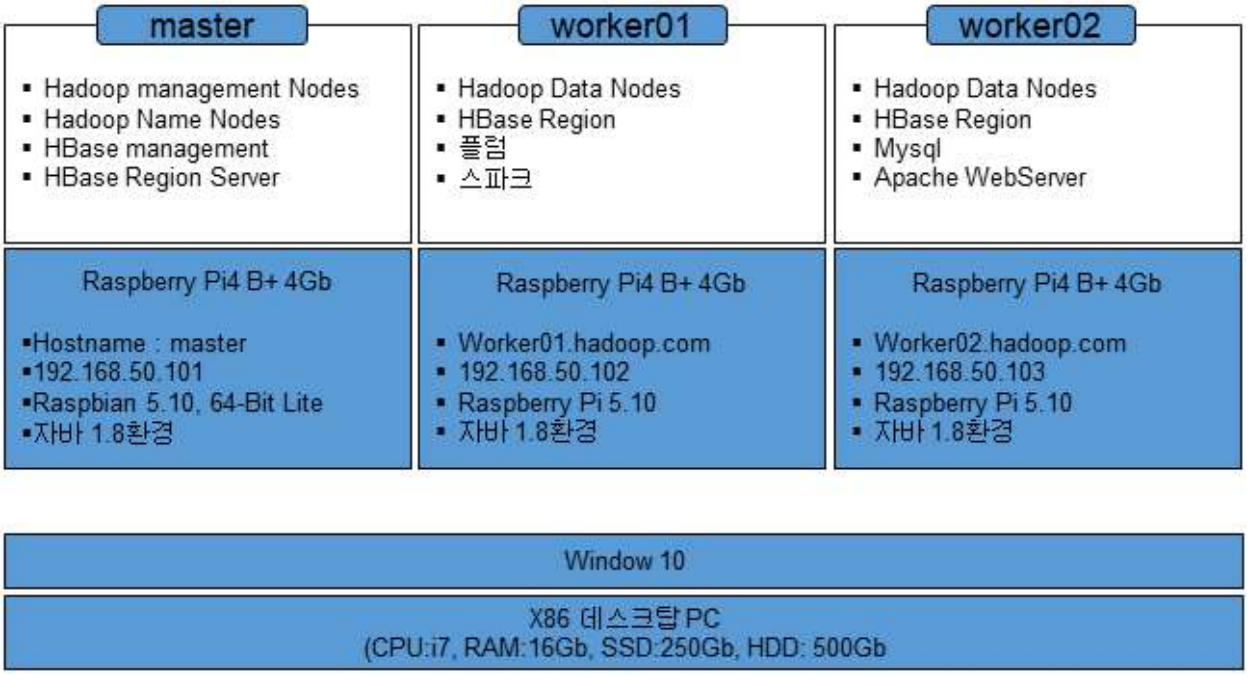
답 :

문제 2-6 (NCS 2.2)	Hadoop의 오픈소스 소프트웨어로 Eco System의 종류를 아는 데로 4가지 이상 기술하시오.

답:

①
②
③
④

Question 3	분산처리 수행 모듈 개발하기



위 그림은 Raspberry Pi를 이용하여 하둡 분산처리시스템을 구성한 구성도이다.
아래 질문에 답하시오.

문제 3-1 (NCS 3.1)	Raspberry Pi에 설치된 하둡 분산 처리 시스템에 접속한 Node에 ①서버의 이름, ②최초 접속한 계정의 이름, ③현재 작업 중인 계정의 이름을 확인 하는 Linux명령는 어떤 것인가?
답: ① ② ③	

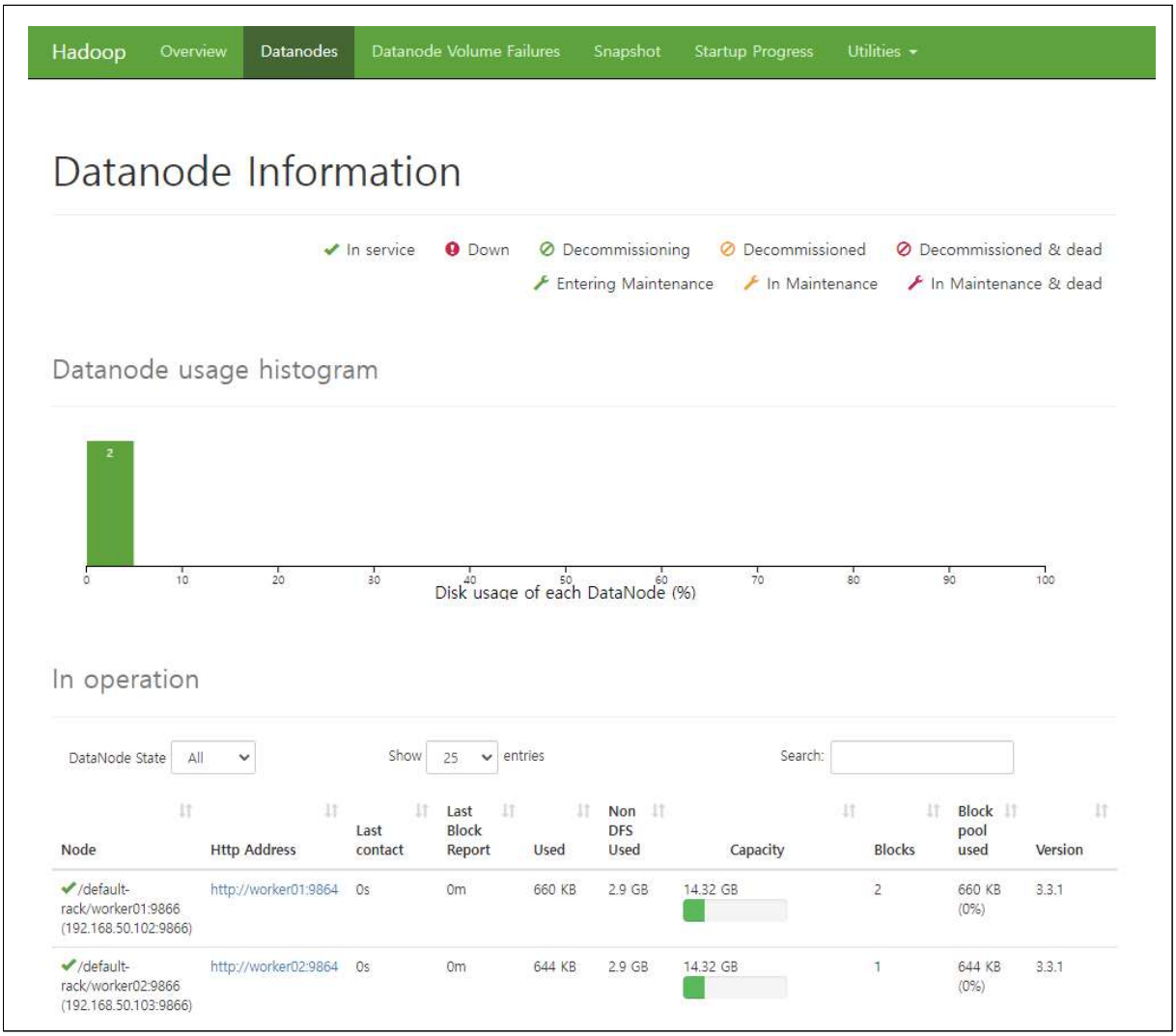
문제 3-2 (NCS 3.1)	Raspberry Pi에 설치된 하둡 분산 처리 시스템에 접속한 Node에 ①IP주소를 확인하고, ②현재 작업 경로위치, ③파일의 목록을 확인 하는 Linux 명령는 어떤 것인가?
답: ① ② ③	

문제 3-3 (NCS 3.1)	Raspberry Pi에 설치된 하둡 분산 처리 시스템에 접속한 Node에 ①hadoop이라는 계정을 만들고, ②/my-hdfs 디렉토리를 만들고, ③ /my-hdfs 디렉토리로 옮겨가는 Linux명령는 어떤 것인가?
답: ① ② ③	

문제 3-4 (NCS 3.3)	Raspberry Pi에 설치된 하둡 분산 처리 시스템을 ①구동시키는 명령어와 실행된 하둡시스템을 ②Web상에서 확인하는 URL주소와 Port번호는 무엇인가?
답: ① ②	

문제 3-5 (NCS 3.2)	Raspberry Pi에 설치된 하둡 분산 처리 시스템을 접속하는 소프트웨어로 File을 전송할 때 사용하는 소프트웨어와, 원격지 Node들에 telnet 접속하여 Node 들을 제어 하는데 사용되는 소프트웨어 이름을 각각 기술하시오.
답: ① ②	

문제 3-6 (NCS 3.3)	아래 그림은 하둡 분산 처리 시스템이 실행된 상태를 조회하는 화면이다. 이 페이지에서 하둡의 ①노드는 몇 개가 구동 되고 있으며, ②각 Node의 IP와 ③Node의 이름은 무엇인가?



답:

①

②

③

문제 3-7 (NCS 3.2)	Raspberry Pi에 설치된 하둡 분산 처리 시스템의 ①root경로의 파일 목록을 보는 명령어와 ②/temp 디렉토리를 만드는 명령어는 무엇인가?

답:

①

②

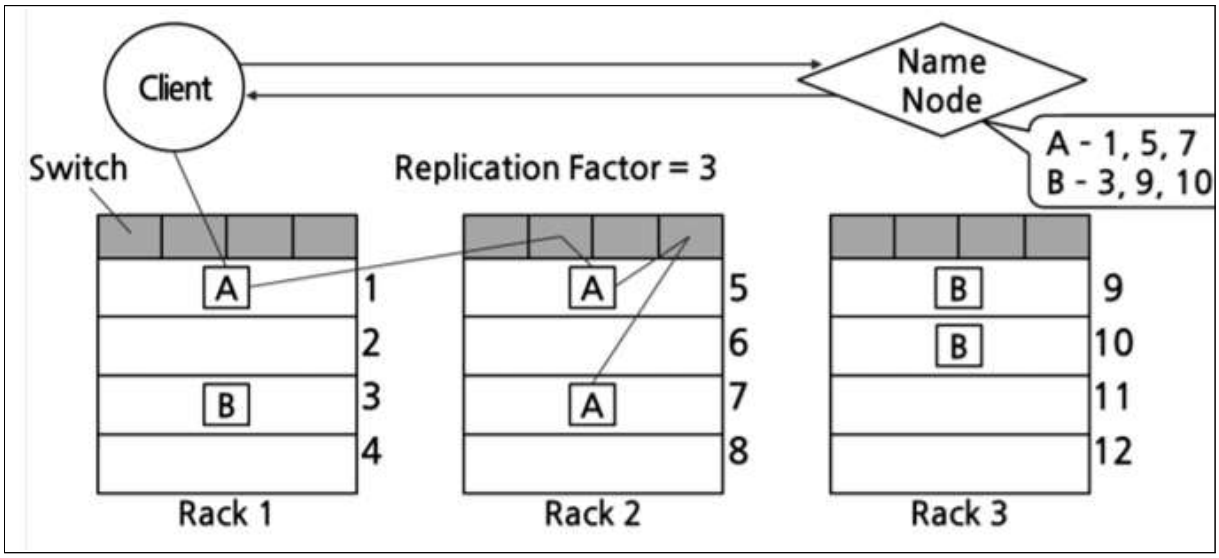
문제 3-8 (NCS 3.2)	Raspberry Pi에 설치된 하둡 분산 처리 시스템에서 현재의 경로에 test.txt파일이 있고 이 파일을 ①/temp directory에 Upload하고, ②Upload한 파일의 내용을 조회하는 명령어는 무엇인가?

답:

①

②

문제 3-9 (NCS 3.2)	아래 그림과 같이 하둡 분산 처리 시스템에 파일을 저장할 때 파일을 여러 개의 블록으로 나누로 나눈 블록을 저장하기 위하여 여러 개의 파일로 복사한다. 하둡 분산처리 시스템은 기본적으로 블록의 크기와 복제 갯수를 정하고 있다. 하둡3.0 이상에서 기본 값으로 정한 블록의 크기와 복제 개수는 몇 개 인지 기술하시오.



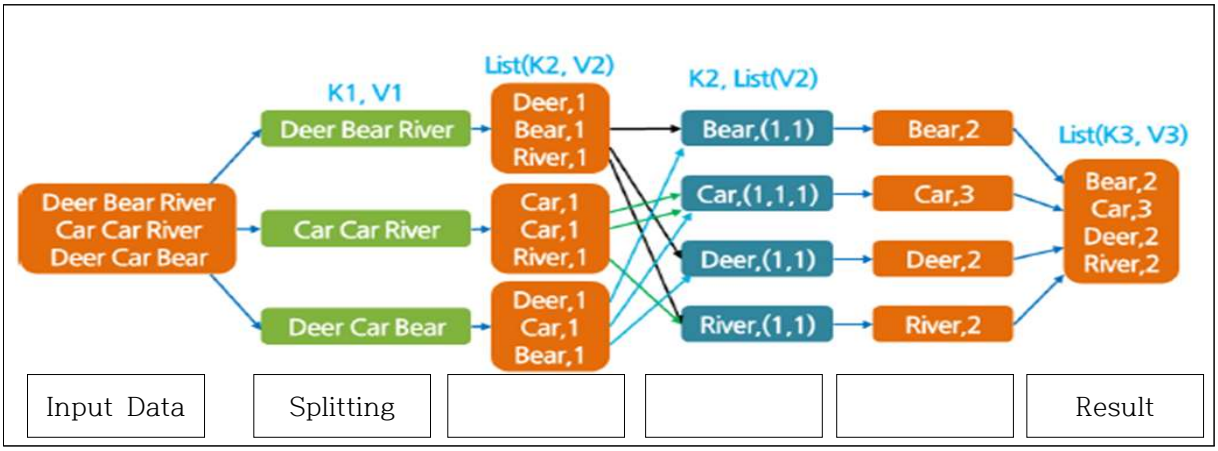
답:

①

②

문제 3-10 (NCS 3.2)	하둡의 분산 처리 시스템을 개발하고 처리를 하기위하여 WordCount 프로그램을 Maven Project를 이용하여 작성하려고 합니다. 메이븐 프로젝트에서 WordCount 작성시 내부에서 사용하는 각종 모듈들을 외부에서 가져와서 사용하게 되는데 이때 Maven Project내에 각종 모듈과 API들을 모아놓고 Maven이 관리하는 파일이 있다. 이 파일의 이름은 무엇인가?
답:	

문제 3-11 (NCS 3.3)	아래 그림은 하둡 분산처리 시스템을 활용하여 단어를 세는 절차를 설명한 그림이다.하둡 분산처리 시스템이 단어를 세는 절차 중에 빈 칸에 들어갈 알맞은 용어는 무엇인지 기술하시오.



답:

①

②

③

문제 3-12 (NCS 3.3)	하둡 분산처리 시스템을 활용하여 단어의 빈도를 세는 WordCount.java 소스 파일을 컴파일하여 wordcount.jar 파일을 작성하였다. 이를 이용하여 하둡 분산처리 시스템에서 단어의 빈도를 세는 프로그램을 실행하여 /tmp/articles.txt파일을 분석하여 결과를 /wc-out에 저장하려고 한다. 이를 실행하기 위한 명령어를 기술하시오.
답:	

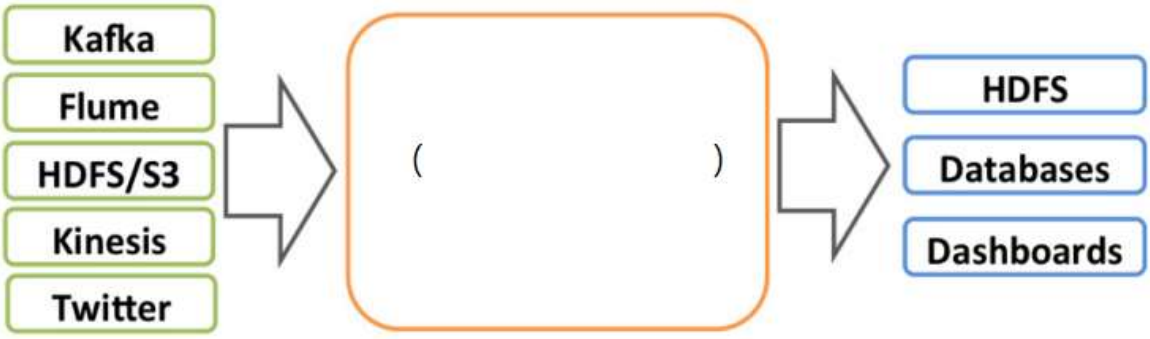
Question 4	실시간 수행 모듈 개발하기

문제 4-1 (NCS 4.1)	고정된 저장소에 저장된 데이터가 아닌 “지속적으로 생성되어 유입되는 데이터”를 말한다. 특히 유입되는 주기가 매우 짧으며, 유입량의 변화량도 큰 특징이 있다. 크리스마스 이브나 세계적인 이슈에 트위터 메시지 전송량이 급증하는 경우를 볼 수 있다. 이러한 스트림 데이터는 IOT 센서, SNS 메신저, 웹서버 로그, 카드 결제 이력 등을 통해 생성되는 데이터이다. 위 에서 서술한 시스템은 무엇을 설명한 것인가?

① 정형데이터　② 비정형 데이터
③ 반정형 데이터　④ 스트리밍 데이터

답 :

문제 4-2 (NCS 4.3)	( )은 다양한 오픈 소스 소프트웨어(Kafka, Flume, Twitter 등)로부터 유입되는 데이터를 추상화된 API를 이용해 복잡한 알고리즘을 쉽게 처리할 수 있으며, 그 결과를 분산 저장 장치에 저장하는 아키텍처를 제공한다. ()에 들어갈 알맞은 오픈 소스 소프트웨어는 무엇인가?



① 하이브(Hive)　② 스파크 스트리밍(Spark Streaming)
③ 임팔라(Implar)　④ 에이치 베이스(HBASE)

답 :

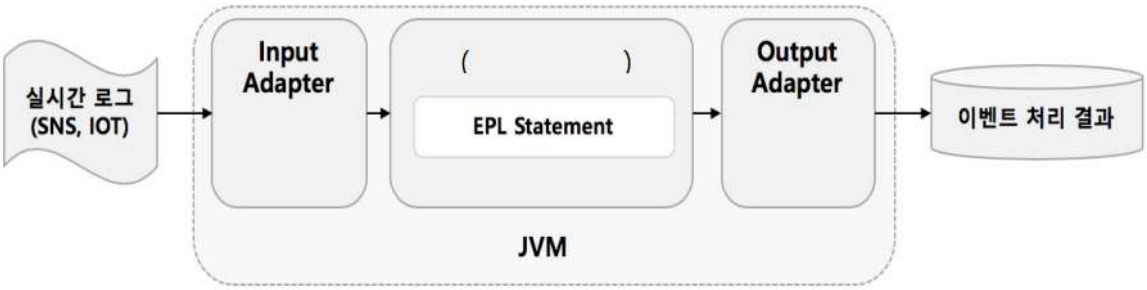
Question 5	이벤트처리 수행 모듈 개발하기

문제 5-1 (NCS 5.1)	( )은 JVM 상에서 동작하는 java 기반의 프로그램으로, 실시간 스트림 데이터의 복잡한 이벤트 처리에 최적화된 오픈 소스 룰 엔진이다. 또한 복잡한 이벤트 처리에 필요한 다양한 조건과 이벤트 발생 조건을 룰로 쉽게 정의할 수 있다. ( )안에 실시간 스트림 데이터를 처리하기 위한 기술은 무엇인가?

① 에스퍼(Esper)	② JBoss Drools
③ 카프카(Kafca)	④ 레디스(Redis)

답 :

문제 5-2 (NCS 5.2)	아래 그림은 에스터 아키텍처 구성도이다. ( )에 들어가는 구성요소는 무엇인가?



답 :