수원대학교 클라우드 기반 데이터레이크 및 분석 중간고사

학번 : _____ 학과 : _____ 이름 : _____

1. 다음 중 DataLake에 대해 올바르게 설명한 것은? **(4)**

   1) Transaction 기반으로 하는 데이터 작업을 말한다.

   2) 현재 업무의 효율적인 처리에만 관심이 있기 때문에 최대 목적은 데이터에 대한 무결성을 확보하는 것이다.

   3) 저장된 데이터에 대해 CRUD를 주로 사용한다.

   4) 여러 가지 다양한 유형의 데이터를 원본 포맷 형태로 저장하는 단일 저장소

2. 다음은 무엇에 대해 설명하고 있는가? **(2)**

   - Data Lake는 이것과 대비되는 개념으로 언급되기도 한다.
   - 특정 팀 또는 사업단위의 요구를 충족시킨다.
   - 규모가 더 작고, 집중적이며 사용자 커뮤니티에 가장 잘 맞는 데이터의 요약본을 말한다.

   1) Data Warehouse 2) Data Mart 3) Database 4) Data Ocean

3. 다음 중 Big Data의 특징으로 볼 수 없는 것은 ? **(4)**

   1) Volume 2) Velocity 3) Variety 4) Variation

4. 다음 중 Data Lake를 선택하는 이유로 보기 어려운 것은?. **(1)**

   1) Database 사용보다 속도가 빠르기 때문에

   2) 기하급수적으로 데이터가 증가하기 때문에

   3) 더 빠른 분석이 필요하기 때문에

   4) 정형데이터 뿐만 아니라 비정형 데이터의 분석이 필요하기 때문에

5. 다음 중 Data Lake 구축 방식 선정에 있어서 잘못된 것은 ? **(2)**

   1) 리소스 소요 변동 폭은 On-Premise보다 Clouse가 높다.

   2) 데이터 민감도는 On-Premise나 Private Cloud 보다 Public Cloud가 높다.

   3) 실시간 처리 필요성은 On-Premise보다 Cloud가 낮다.

   4) 보유하고 있는 리소스(인력/비용) 수준은 On-Premise나 Private Cloud보다 Public Cloud가 낮다.

6. 다음은 Data Lake 구현 단계 중 **설계 단계**에 대한 작업을 정리하였다. **설계 단계**에 잘 못 포함된 작업은 무엇인가? **(2)**

   1) 구성 요소별 필요 기능 정의

   2) 구축 방식 선정

   3) H/W 아키텍처 설계

   4) 참조 모델 기반 Layer 구분

7. 다음 중 Data Lake 구현 방식을 On-Premise와 Cloud로 나눠서 비교했을 때 Cloud로 구현시 유리하지 않는 항목은 무엇인가? **(3)**

   1) 리소스 소요 변동 폭이 클 경우에 유리하다.

   2) 배치성 업무 처리 비중이 높을 시 유리하다

   3) 전사 공통 비용 기반으로 운영 시 유리하다.

   4) 향후 확장 가능성이 높을 경우 유리하다.

8. 다음 중 Apache Spark에 대한 설명으로 잘못된 것은? **(4)**

   1) 데이터 처리 분야에서 가장 규모가 큰 프로젝트이다.

   2) 인메모리 기반의 데이터 처리는 하둡 대비 100배 빠르다.

   3) 주요 아키텍쳐는 Batch/Streaming Data 처리, SQL 분석, 구조화된 데이터 처리, 머신 러닝 및 그래프 계산이다.

   4) 빅데이터 워크로드에 사용하는 상용 통합 분석 엔진이다.

9. 다음 AWS 서비스 중 분석 사용자가 여러 소스의 데이터를 쉽게 검색, 준비, 이동, 통합할 수 있도록 하는 서버리스 데이터 통합 서비스는 무엇인가? **(4)**

   1) AWS RDS　　2) Amazon Athena　　3) Amazon S3　　4) AWS Glue

10. 다음 중 aws configure 설정 시 필요하지 않는 항목은? **(1)**

1) Default language
2) AWS Access Key ID
3) Default output format
4) AWS Secret Access Key

11. 다음 중 AWS CLI를 이용하여 본인의 S3 Bucket의 내용물(객체)을 확인하기 위한 명령은 무엇인가? **(4)**

1) $ aws s3 cp s3://{Bucket Name}/
2) $ aws git ls {Bucket Name}/
3) $ aws s3 ls http://{Bucket Name}/
4) $ aws s3 ls s3://{Bucket Name}/

12. 다음 중 Apache Spark를 위한 환경설정 중 반드시 설정할 필요가 없는 환경변수는? **(2)**

1) JAVA_HOME
2) TEMP
3) SPARK_HOME
4) HADOOP_HOME

13. 다음은 Apache Spark의 어떤 개념을 설명하고 있는가? **(3)**

- Apache Spark의 RDD나 Dataset를 구성하고 있는 최소 단위 객체
- Spark의 성능과 리소스 점유량을 크게 좌우할 수 있는 가장 기본적인 개념
- 1 Core = 1 Task = 1 ( )
- 결국, 이것의 크기가 Core 당 필요한 메모리 크기를 결정한다.

1) Data Frame
2) Data Catalog
3) Partition
4) ETL

14. 다음 중 AWS의 최소 권한 원칙을 가장 잘 설명한 것은 무엇인가? **(3)**

1) IAM 사용자를 하나 이상의 IAM 그룹에 추가
2) 액세스 제어 목록에서 패킷의 권한을 확인
3) 특정 작업을 수행하는 데 필요한 권한만 부여
4) 하나 이상의 디바이스에서 시작하는 서비스 거부 공격을 수행

15. 다음 AWS 서비스 중 Data Lake의 저장소 역할을 담당하는 서비스는 무엇인가? **(3)**

1) AWS Glue 2) Amazon Athena 3) Amazon S3 4) CloudWatch

16. 다음 AWS 서비스 중 AWS Glue Studio에서 Job을 실행 중 오류가 발생하면, 어느 서비스에서 해당 오류를 확인할 수 있는가? **(4)**

1) AWS Glue Workflow 2) Amazon Athena의 설정 탭
3) Amazon S3 대시보드 4) CloudWatch Error Log

17. 다음 설명은 AWS의 어떤 서비스를 설명하고 있는가? **(1)**

- PiB 규모의 데이터에 대해 표준 SQL문에 기반한 질의를 수행할 수 있다.
- 데이터 소스에 대응하는 테이블 메타 정보만 생성하면 바로 쿼리를 수행할 수 있으며, 쿼리 수행 속도 또한 매우 빠르다.
- log, json, csv 등 다양한 파일을 읽고 표준 SQL을 사용하여 질의할 수 있다.
- AWS Glue의 Crawler를 통해 수집한 데이터의 내용을 질의를 통해 확인할 수 있다.

1) Amazon Athena 2) AWS EMR 3) Parameter Store 4) KMS

18. 다음 중 DataLake이름으로 Python 가상 환경 구축시 잘못된 명령은? **(4)**

1) python -m venv DataLake
2) $ source DataLake/bin/activate (Linux, macOS)
3) virtualenv DataLake
4) env --virtual DataLake

19. 다음 중 Pyspark 코드를 통해 Amazon S3에 접근하기 위한 경로로 올바르게 작성한 것은? 단, Bucket의 이름은 datalake-bucket이다. **(3)**

1) path = 'http://datalake-bucket'
2) path = 's3://datalake-bucket'
3) path = 's3a://datalake-bucket'
4) path = 'https://s3.datalake-bucket'

20. 다음 중 AWS Glue Crawler의 schedule에서 기본적으로 지원하지 않는 항목은? **(3)**

1) On demand
2) Hourly
3) Secondly
4) Monthly

21. 다음 중 AWS Glue Studio의 Script를 실행할 때 제일 마지막에 있는 코드로서 Job의 결과를 반영하고 Script가 모두 마치면 수명 주기를 수동으로 종료해야 하는 데 이 때 사용하는 코드는 무엇인가? **(2)**

1) job.end()
2) job.commit()
3) job.finish()
4) job.finalize()

22. 다음 중 Hadoop Ecosystem에 포함되지 않는 서비스는? **(3)**

1) Apache Kafka
2) Apache Spark
3) MySQL
4) Apache HIVE

23. 다음 중 Data Engineer에 대해 잘 못 설명한 것은? **(3)**

1) 데이터 파이프라인 및 데이터 처리 시스템을 설계, 구현 및 운영
2) Data warehousing, 데이터 통합 및 모델링에 대한 이해
3) 통계 기술과 머신러닝 모델을 사용하여 예측하고 추세를 분석
4) SQL, NoSQL Database, Python, Java, Scala

24. 다음 중 비정형 데이터로 적합하지 않는 포맷은? **(1)**

1) JSON 2) Video 3) Image 4) Audio

25. 다음 중 Data Lake를 통해 수행할 수 없는 작업은? **(5)**

1) 데이터의 수집과 저장
2) 데이터의 보안 및 보호
3) 데이터 분석 및 통찰력
4) 카탈로그 작성 및 검색
5) 데이터를 통한 통제 및 감시