

데린이를 위한 꿀 팁

@Youngick Kim

2023-09-07



데린이들의 고민

- 데이터 관련 직무 신입 또는 이직을 원하는데 ...
- 뭘 공부하면 좋을까요?
- 자격증 따면 도움이 되나요?

21일 전

엑셀로 다운받고... 한글은 깨지네요 ㅎㅎ;;

image.png ▼

Name	User ID	Title	대행	RSVP	RSVPed on	Joined Group	URL of Member
Kim	user 229959622		Yes	Yes	5,000 2023 8 12		
Kim	user 1944614	Co-Organizer	Yes	Yes	2023 7 27		
	user 4186488		Yes	Yes	5,000 2023 8 11		
Yu	user 2352027	Co-Organizer	Yes	Yes	5,000 2023 7 27		
ak	user 56228942		Yes	Yes	5,000 2023 8 10		
	user 398202560		Yes	Yes	5,000 2023 8 15	2023 7 23	https://www.meetup.com/ko-KR/...
eon	user 229795935		Yes	Yes	5,000 2023 8 9	2023 8 9	https://www.meetup.com/ko-KR/...
Yoonian	Co-Organizer	Yes	Yes	Yes	2023 7 27	2018 5 9	https://www.meetup.com/ko-KR/...
Lee	user 208734007		Yes	Yes	5,000 2023 8 2	2022 2 9	https://www.meetup.com/ko-KR/...
	user 400148773		Yes	Yes	5,000 2023 8 17	2023 8 17	https://www.meetup.com/ko-KR/...
	user 214300420		Yes	Yes	2023 8 2	2016 10 4	https://www.meetup.com/ko-KR/...
	user 260971508		Yes	Yes	5,000 2023 8 14	2022 7 3	https://www.meetup.com/ko-KR/...
G. will2you		Yes	Yes	Yes	5,000 2023 7 27	2016 10 4	https://www.meetup.com/ko-KR/...
	user 269275641		Yes	Yes	5,000 2023 8 10	2022 2 16	https://www.meetup.com/ko-KR/...
	user 328168055		Yes	Yes	5,000 2023 7 27	2021 4 2	https://www.meetup.com/ko-KR/...

한글이 깨져요

```
8 JunSeong user 229795935 Yes W5,000 2023년 8월 9일 오전 11:46 2023년 8월 9일
9 Ho Yoon yoonian Co-Organizer Yes Yes 2023년 7월 27일 오후 5:05 2018년 5월 9일
10 eon Lee user 208734007 Yes W5,000 2023년 8월 2일 오전 9:52 2022년 2월 9일
11 user 400148773 Yes W5,000 2023년 8월 17일 오후 1:04 2023년 8월 17일 https://www.
12 user 214300420 Yes 2023년 8월 2일 오후 4:52 2016년 10월 4일 https://www.
13 user 260971508 Yes W5,000 2023년 8월 14일 오후 10:39 2022년 7월 3일 http
14 / M.G. Chang will2you Yes Yes W5,000 2023년 7월 27일 오후 8:55 2016년 10월 4
15 user 269275641 Yes W5,000 2023년 8월 10일 오후 2:51 2022년 2월 16일 https://
16 user 328168055 Yes W5,000 2023년 7월 27일 오후 9:56 2021년 4월 2일 https://
```

21일 전

저도 Numbers로 여니깐 안 깨졌어요. 엑셀이 엑셀파일을 제대로 못 보여주네요 ㅠ

1

AWSKRUG_ _#gametech_ _-_8 _30 ().xls

인코딩(N) 언어(L) 설정(T) 도구

ANSI

• UTF-8

UTF-8 BOM

UCS-2 BE

UCS-2 LE

문자 집합

ANSI로 변환

UTF-8로 변환

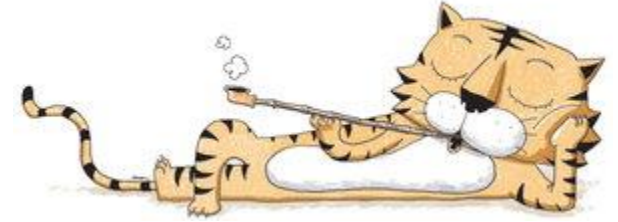
UTF-8 BOM으로 변환

UCS-2 BE로 변환

UCS-2 LE로 변환

1	Name	User ID	Title	이벤트 조직자	RSVP	게스트	RSVPed o	
2	Don	Lim user	233210132		Yes	2023년 7월 28일		Me
3	Hyo	Lim user	310321586		Yes	2023년 8월 6일		tps
4	Hyo	Kim user	194461433	Co-Organizer	Yes	Yes		h
5	JEO	Shik user	70909562		Yes	2023년 8월		16년
6	Sae	Black user	321922804	Co-Organizer	Yes	Yes		4일h
7	Som	user	186522266		Yes	2023년 8월 4일		20년
8	Tae	Park user	184515293	Co-Organizer	Yes	Yes	2023년 7월 28일 오후 3:25	2
9	Yoh	user	365607985	Yes		2023년 8월 9일 오후 9:28	2022년 12월 8일	https
10	you	kim user	246148128	Co-Organizer	Yes	Yes	2023년 8월 1일 오후 10:13	2
11	길	은user	385163093	Yes		2023년 8월 24일 오후 7:37	2023년 8월 24일	https://w
12	김	주user	359118423	Yes		2023년 8월 14일 오전 10:14	2022년 4월 5일	https://www.m
13	김	호user	377444683	Yes		2023년 7월 28일 오후 6:23	2022년 12월 1일	https://w
14	노	정user	395452940	Yes		2023년 7월 31일 오후 5:45	2023년 6월 21일	https://w
15	이	주user	214300420	Yes		2023년 8월 2일 오후 4:53	2016년 10월 4일	https://w
16	재	오 user	379941861	Yes		2023년 7월 31일 오후 5:37	2023년 1월 25일	https
17	정	원user	361619956	Yes		2023년 8월 21일 오전 8:19	2022년 5월 10일	https://w
18								

호랭이 담배 피던 시절...



- 10년전 빅 데이터(하둡) 해보려고 이직
<https://www.databricks.com/kr/glossary/hadoop>

하둡이란 무엇입니까?

"하둡"이란 무엇을 의미할까요? 더 중요한 것은, "하둡"은 무엇의 약자일까요? 사실, 고가용성 분산형 객체 지향적 플랫폼(High Availability Distributed Object Oriented Platform)을 뜻합니다. 하둡 기술은 바로 이런 장점을 개발자에게 제공합니다. 즉, 객체 지향적 작업을 병렬 분산하여 고가용성을 확보할 수 있습니다.

Apache Hadoop은 오픈 소스, Java 기반 소프트웨어 플랫폼으로 빅데이터 애플리케이션용 데이터 처리와 스토리지를 관리하는 역할을 합니다. 하둡 플랫폼은 컴퓨터 클러스터 내 여러 노드에 걸쳐 하둡 빅데이터와 분석 작업을 분배하며, 그 과정에서 작업을 병렬식으로 실행 가능한 작은 크기의 워크로드로 분해합니다.

하둡은 구조적, 비구조적 데이터를 처리할 수 있으며 단 한 대의 서버에서 시스템 수천 대 규모로 안정적으로 확장합니다.



하둡 프로그래밍이란 무엇인가요?

하둡 프레임워크에서 코드는 대부분 Java로 작성되지만, 일부 네이티브 코드는 C를 기반으로 합니다. 또한, 명령줄 유틸리티는 셸 스크립트로 작성되는 것이 일반적입니다. 하둡 MapReduce의 경우, Java를 가장 흔히 사용하지만 사용자는 하둡 스트리밍과 같은 모듈을 통해 원하는 프로그래밍 언어로 맵을 구현하고 함수를 줄일 수 있습니다.

하둡 데이터베이스란 무엇인가요?

하둡은 데이터 스토리지나 관계형 데이터베이스를 위한 솔루션이 아닙니다. 대신 오픈 소스 프레임워크로써, 실시간으로 엄청난 양의 데이터를 동시에 처리하는 것을 목적으로 합니다.

데이터는 HDFS에 저장됩니다. 그러나 비구조적이어서 관계형 데이터베이스로 간주하기 어렵습니다. 사실, 하둡을 사용하면 데이터를 비구조적, 반구조적, 구조적 형식으로 저장할 수 있습니다. 기업에서는 비즈니스 요구 사항을 충족하고 그 이상까지 기대할 수 있는 방식으로 더욱 유연하게 빅데이터를 처리할 수 있습니다.

하둡은 언제 발명되었나요?

Apache Hadoop은 Yahoo나 Google과 같은 검색 엔진이 막 출발한 시점에서 끊임없이 늘어나는 빅데이터를 처리하고 웹 결과를 더 빨리 제공해야 한다는 필요성이 절실해지면서 탄생했습니다.

당시 Google의 **MapReduce**라는 프로그래밍 모델이 있었는데, 이 모델은 하나의 애플리케이션을 여러 부분으로 나누어 서로 다른 여러 노드에서 실행되도록 하는 방식을 취했습니다. 여기에서 아이디어를 얻은 Doug Cutting과 Mike Cafarella가 2002년에 Apache Nutch 프로젝트를 진행하던 중 하둡을 시작하게 됐습니다. 뉴욕 타임스 기사에 따르면 Hadoop이라는 이름은 Doug이 아들의 장난감 코끼리 이름에서 따온 것이라고 합니다.

몇 년이 지난 뒤 하둡은 Nutch에서 분할되어 나왔습니다. Nutch는 웹 크롤러 요소에 집중했지만, 하둡은 분산형 컴퓨팅 및 처리 부분을 담당하게 된 것입니다. Cutting이 Yahoo에 입사하고 2년 후인 2008년, Yahoo에서 하둡을 오픈 소스 프로젝트로 릴리스하였습니다. 2012년 11월에 Apache Software Foundation(ASF)에서 하둡을 Apache Hadoop이라는 이름으로 일반 대중에게 제공하게 되었습니다.

하둡은 어떤 유형의 데이터베이스인가요?

기술적으로 하둡은 SQL이나 RDBMS 등의 데이터베이스와는 다릅니다. 대신, 하둡 프레임워크는 사용자에게 다양한 데이터베이스 유형에 대한 처리 솔루션을 제공합니다.

하둡은 기업이 단시간에 방대한 데이터를 처리하도록 도와주는 소프트웨어 생태 시스템입니다. 이런 작업은 대규모 병렬 컴퓨터 처리를 활용하기 때문에 가능합니다. Apache HBase와 같은 여러 데이터베이스는 수백, 수천 개의 상용 서버에 저장된 데이터 노드 클러스터에 흩어져 있습니다.

하둡은 어떤 언어로 작성되었나요?

하둡 프레임워크 자체는 대부분 Java를 기반으로 합니다. 다른 프로그래밍 언어로는 C로 작성한 네이티브 코드와 명령줄용 셸 스크립트를 사용합니다. 그러나 하둡 프로그램은 Python, C++ 등의 다양한 프로그래밍 언어로 작성할 수 있습니다. 따라서 프로그래머는 자신에게 익숙한 도구를 사용해서 유연하게 일할 수 있습니다.

하둡은 어떤 영향을 미쳤나요?

하둡은 빅데이터 분야에서 중대한 발전이었습니다. 사실, 하둡은 현행 클라우드 데이터 레이크의 기초 토대로 인정받고 있습니다. Hadoop은 컴퓨팅 파워를 민주화하여 기업에서 무료, 오픈 소스 소프트웨어와 저렴한 상용 하드웨어를 사용해 확장할 수 있는 방식으로 빅데이터 세트를 분석, 쿼리할 수 있게 해준 역할을 했습니다.

이것이 중대한 의의가 있는 이유는 하둡 덕분에 당시까지 대세였던 상용 (proprietary) 데이터 웨어하우스(DW) 솔루션과 폐쇄형 데이터 형식에 실질적인 대안을 제시해주었기 때문입니다.

하둡이 도입되면서 기업에서 엄청난 양의 데이터를 저장, 처리할 능력을 신속히 확보할 수 있게 되었고, 컴퓨팅 파워를 증가하고 내결함성, 데이터 관리 유연성, DW에 비해 저렴한 비용은 물론 뛰어난 확장성까지 얻게 되었습니다. 궁극적으로 Hadoop은 빅데이터 분석 분야의 향후 개발을 위해 길을 개척했다고 볼 수 있습니다. Apache Spark가 가장 대표적인 예입니다.

SQL on Hadoop



Apache Hive

Apache Hive는 초기에 하둡으로 SQL을 쿼리하는 데 일반적으로 사용했던 솔루션이었습니다. 이 모듈은 MySQL의 동작, 구문, 인터페이스를 에뮬레이션하여 프로그래밍을 단순화합니다. Java API와 JDBC 드라이버가 내장되어 있기 때문에 Java 애플리케이션을 많이 사용한다면 좋은 옵션이 됩니다. Hive는 개발자에게 빠르고 간단한 솔루션을 제공하지만 다소 느리고 읽기 전용 기능만 제공하기 때문에 상당히 제한적입니다.

다른 기술들

- No SQL 은 어디로?
- RDBMS 는 이제 사용 안하나요?



≡ 은빛 총알은 없다

文A 8개 언어 ▼

문서 토론

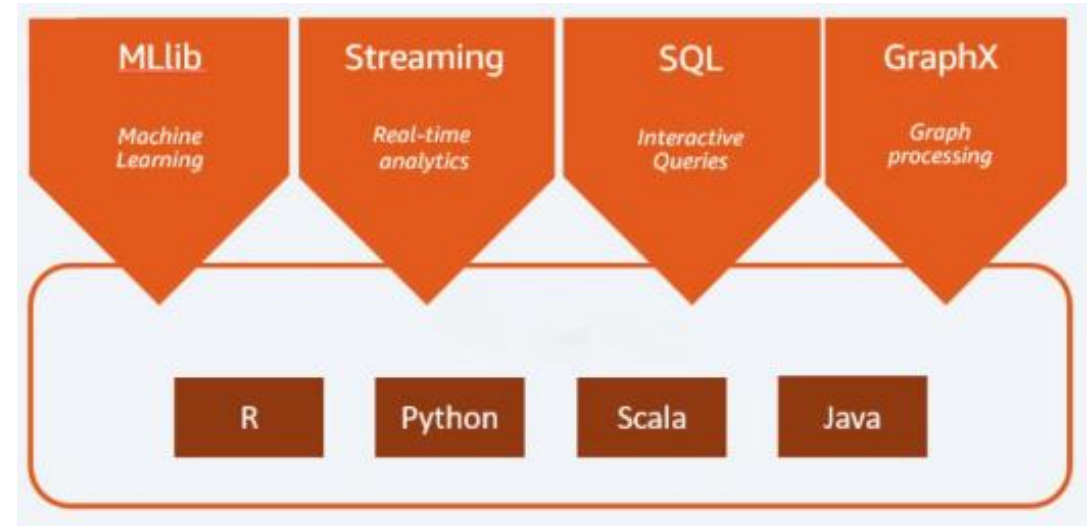
도구 ▼

위키백과, 우리 모두의 백과사전.

은빛 총알은 없다(No Silver Bullet - Essence and Accident in Software Engineering, 은환탄은 없다, 은탄이란 없다)는 1986년 튜링상 수상자 [프레드 브룩스](#)가 쓴 소프트웨어 엔지니어링에 관해 널리 논의된 논문이다.^{[1][2]} 브룩스는 "기술이든 관리 기법이든 한쪽으로만 이루어진 개발은 없으며 그 자체로 10년 안에 생산성, 신뢰성, 단순성 면에서 [크기](#) 정도의 개선만을 약속한다."고 논한다.



Spark 등장



Apache Spark는 빅데이터 워크로드에 쓰이는 오픈 소스 분석 엔진입니다. 배치는 물론 실시간 분석과 데이터 처리 워크로드도 처리할 수 있습니다. Apache Spark는 2009년 캘리포니아 대학교 버클리 캠퍼스에서 연구 프로젝트로 시작되었습니다. 연구진은 하둡 시스템에서 처리 작업의 속도를 높일 방법을 강구하고 있었습니다. 이 엔진은 하둡 MapReduce 기반이었으며 MapReduce 모델을 확장하여 더 많은 연산 유형에 이를 효율적으로 이용하고자 하였는데, 인터랙티브 쿼리와 스트림 처리 등이 대표적인 예입니다. Spark는 Java, Scala, Python과 R 프로그래밍 언어에 네이티브 바인딩을 제공합니

- <https://www.databricks.com/kr/glossary/what-is-apache-spark>



databricks

Apache Spark as a Service란 무엇 입니까?

Apache Spark는 고속 실시간 대규모 데이터 처리를 위한 오픈 소스 클러스터 컴퓨팅 프레임워크입니다. Spark는 2009년 UC 버클리 AMPLab에서 탄생한 이래 큰 성장을 이루었습니다. 지금은 빅데이터 부문에서 가장 큰 오픈 소스 커뮤니티로 평가되며 50여 개 조직과 단체에서 200여 명이 기여하고 있습니다. Databricks는 자사 Apache Spark 최적화 버전을 여러 가지 클라우드에서 Spark-as-a-Service로 호스팅합니다. 여기에 일련의 기본 내장 애플리케이션이 함께 제공되어 데이터에 액세스, 이를 분석하는 속도가 한층 빨라집니다. Spark as a Service는 빅데이터에 작용하는 Spark의 무수히 많은 기능을 활용합니다. 예를 들어 스트리밍 데이터를 다루거나 그래프 연산을 수행

- <https://www.databricks.com/kr/glossary/what-is-apache-spark>

데이터 레이크하우스란 무엇입니까?

데이터 레이크하우스는 데이터 레이크가 가지고 있는 유연성, 비용 효율성, 그리고 대용량 지원 기능에 더해, 데이터 웨어하우스의 데이터 관리 기능과 ACID 트랜잭션을 통합한 새로운 형태의 오픈 데이터 관리 아키텍처로, 모든 데이터를 대상으로 비즈니스 인텔리전스(BI)와 머신 러닝(ML)을 지원합니다.

데이터 레이크하우스: 단순함, 유연함 그리고 저렴한 비용

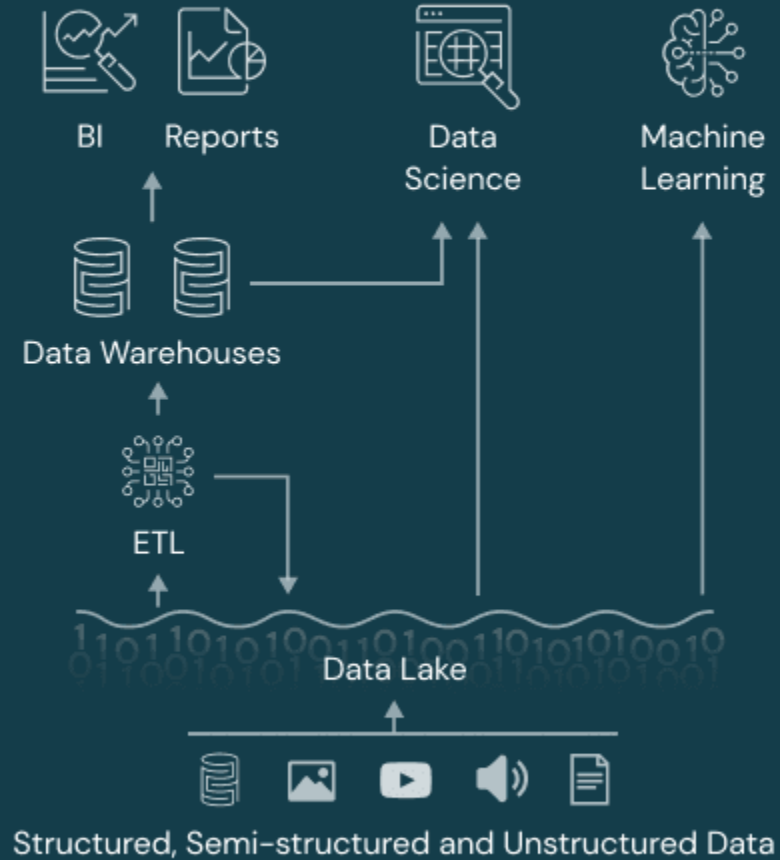
데이터 레이크하우스는 새로운 오픈 시스템 디자인입니다. 데이터 웨어하우스와 비슷한 데이터 구조와 데이터 관리 기능을 구현하되, 데이터 레이크에 쓰이는 저가 스토리지 상에 직접 구현하였습니다. 이 두가지를 하나로 병합함으로써 데이터 팀의 작업 속도가 빨라지는데, 이는 여러 시스템에 액세스하지 않고도 데이터를 사용할 수 있기 때문입니다. 또한 데이터 레이크하우스를 사용하면 팀원들이 가장 완전한 버전의 최신 데이터를 이용하여 데이터 사이언스, 머신 러닝과 비즈니스 분석 프로젝트를 수행할 수 있습니다.

- <https://www.databricks.com/kr/glossary/data-lakehouse>

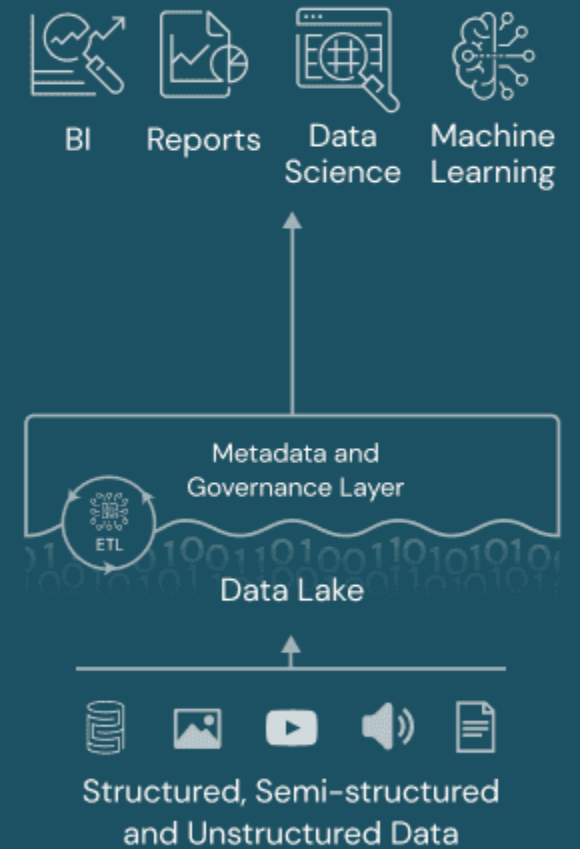
Data Warehouse



Data Lake

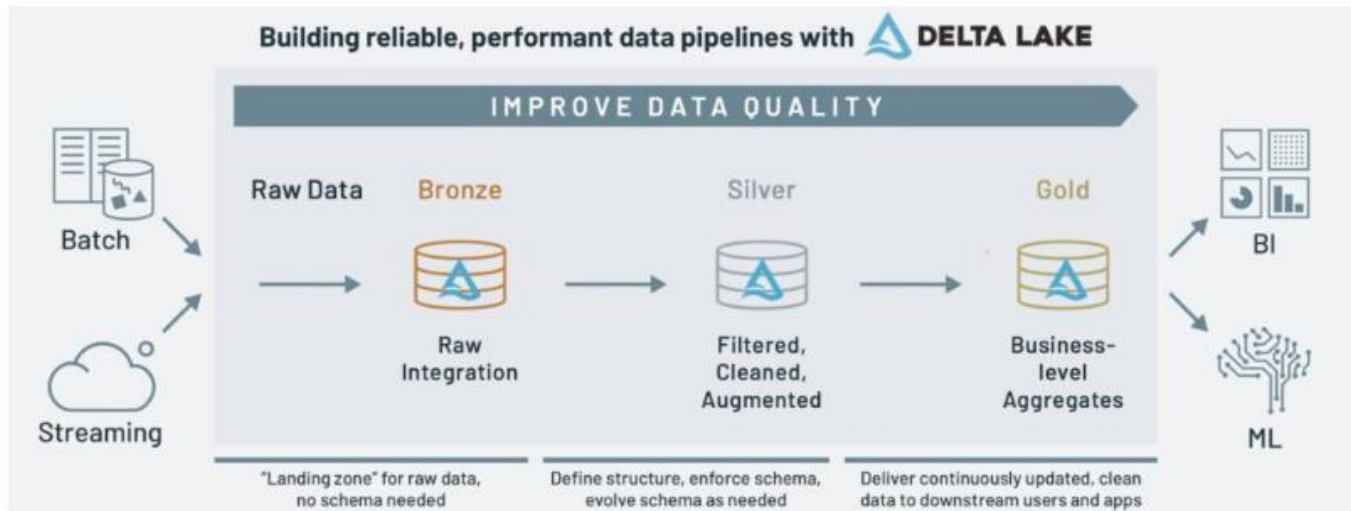


Data Lakehouse



Medallion 아키텍처란 무엇입니까?

메달리온 아키텍처는 레이크하우스에 논리적으로 데이터를 정리하는 데 사용하는 데이터 설계 패턴입니다. 이 아키텍처의 목표는 데이터가 아키텍처의 각 레이어를 통과하는 동안(브론즈 → 실버 → 골드 레이어 테이블) 데이터의 구조와 품질을 증분적, 점진적으로 개선하는 것입니다. 메달리온 아키텍처는 "멀티 홉" 아키텍처라고 부르기도 합니다.



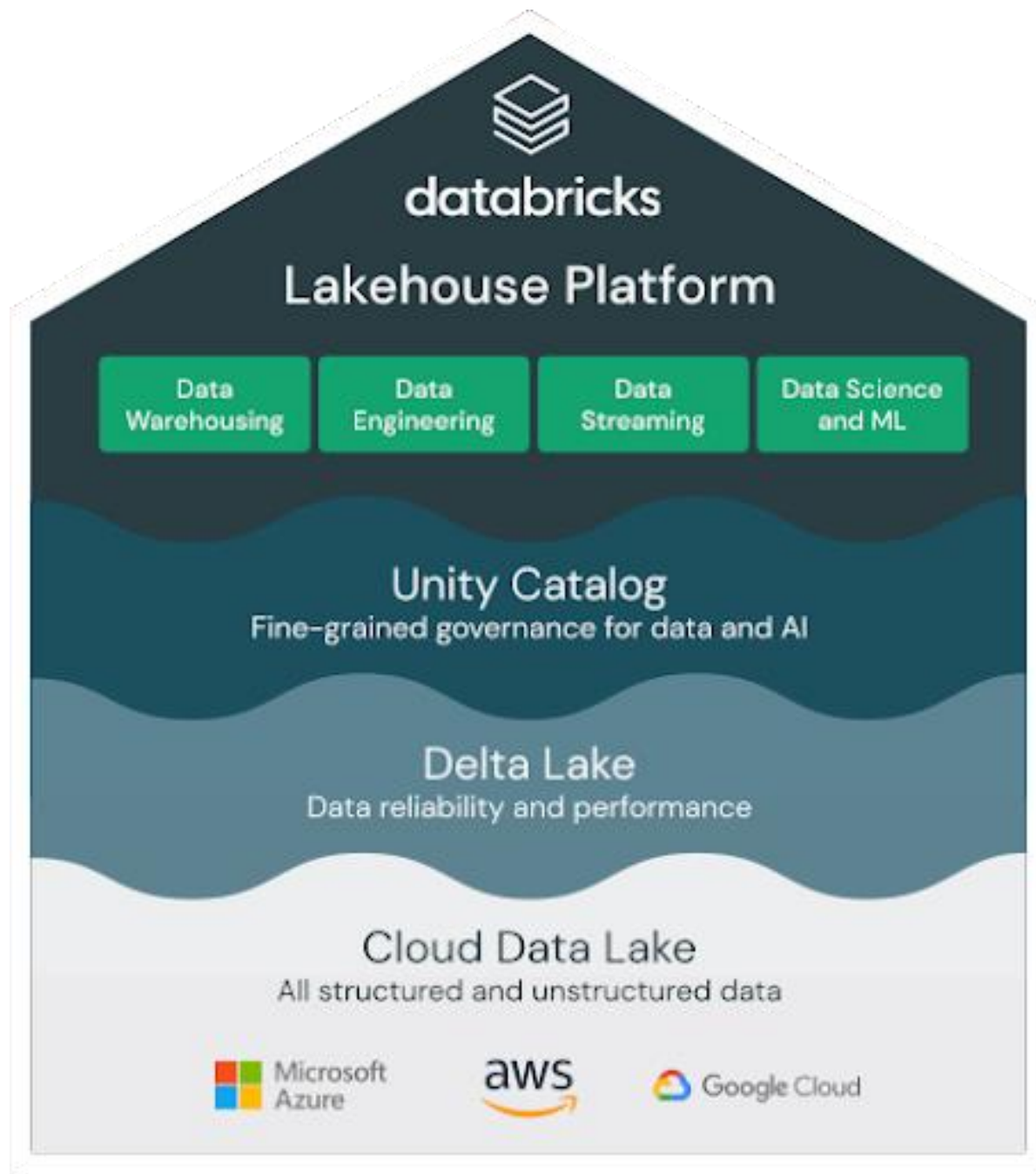
- <https://www.databricks.com/kr/glossary/medallion-architecture>

레이크하우스 아키텍처의 장점

- 간단한 데이터 모델
- 쉽게 이해하고 구현 가능
- 증분적 ETL 지원
- 언제든지 원시 데이터에서 테이블 생성 가능
- ACID 트랜잭션, 시간 이동

레이크하우스에 대한 간단한 설명

레이크하우스는 데이터 레이크와 데이터 웨어하우스의 장점만을 결합한 새로운 데이터 플랫폼 아키텍처 패러다임입니다. 현대적 레이크하우스는 매우 확장성이 높고 성능이 우수한 데이터 플랫폼으로, 원시 데이터 세트와 준비된 데이터 세트를 모두 호스팅하여 기업에서 빠르게 사용할 수 있도록 지원합니다. 또한, 고급 비즈니스 인사이트를 확보하고 결정에 도움을 받을 수 있습니다. 데이터 사일로를 무너트리고, 하나의 플랫폼에서 회사 전체의 권한이 있는 사용자에게 매끄럽고 안전한 데이터 액세스를 제공합니다.



브론즈 레이어(원시 데이터)

브론즈 레이어에는 외부 소스 시스템의 모든 데이터가 들어갑니다. 이 레이어의 테이블 구조는 소스 시스템 테이블 구조에 "그대로" 대응하며, 로드 날짜/시간, 프로세스 ID 등을 캐캡처하는 메타데이터 컬럼이 추가됩니다. 이 레이어는 변경 데이터를 빠르게 캡처할 뿐만 아니라, 소스(콜드 스토리지)의 과거 아카이브, 데이터 리니지, 감사 기능, 필요할 경우 소스 시스템에서 데이터를 다시 읽지 않고도 재처리하는 기능을 제공하는 것이 핵심입니다.

실버 레이어(정리와 순응이 끝난 데이터)

레이크하우스의 **실버 레이어**에서는 브론즈 레이어의 데이터에 매칭, 병합, 순응, ("적당한 수준"으로) 정리를 적용합니다. 실버 레이어에서는 모든 주요 비즈니스 단체, 개념, 트랜잭션에 대한 "엔터프라이즈 뷰"를 제공합니다. (예: 마스터 고객, 스토어, 중복이 없는 트랜잭션, 교차 참조 테이블).

실버 레이어는 다른 소스의 데이터를 엔터프라이즈 뷰로 가져오고, 즉석 보고를 위한 셀프 서비스 분석과 고급 분석, ML을 지원합니다. 실버 레이어는 부서 애널리스트, 데이터 엔지니어, 데이터 사이언티스트에게는 소스 역할을 하면서, 이들이 프로젝트와 분석을 추가로 생성하여 골드 레이어에 있는 회사 및 부서 데이터 프로젝트를 통해 비즈니스 문제에 답할 수 있도록 돕습니다.

골드 레이어(큐레이션된 비즈니스 레벨 테이블)

일반적으로 레이크하우스 **골드 레이어**에 있는 데이터는 바로 사용할 수 있는 "프로젝트 별" 데이터베이스에 정리됩니다. 골드 레이어는 보고용으로 사용하고, 조인의 개수가 적고 더욱 비정규화된 읽기 최적화 데이터 모델을 사용합니다. 여기에 데이터 변환과 데이터 품질 규칙의 마지막 레이어가 적용됩니다. 고객 분석, 제품 품질 분석, 재고 분석, 고객 세그먼테이션, 제품 추천, 마케팅/영업 분석 등의 프로젝트에서 마지막 표시 레이어가 여기에 들어갑니다. 레이크하우스의 골드 레이어에는 주로 Kimball 스타일 스타 스키마 기반 데이터 모델이나 Inmon 스타일 데이터 마트가 들어가는 사례가 많습니다.

CSV

TSV

JSON



ICEBERG



Parquet



Apache
orc™

CSV

TSV

JSON



Parquet

ICEBERG



Apache
orc™

Parquet란 무엇입니까?

Apache Parquet는 효율적인 데이터 스토리지와 검색을 지원하도록 설계되었으며, 컬럼 중심의 오픈 소스 데이터 파일 형식입니다. 복잡한 데이터를 일괄적으로 처리하는 기능을 더욱 향상하여 효율적인 데이터 압축 및 인코딩 방식을 제공합니다. Apache Parquet는 배치 및 인터랙티브 워크로드에 공통적인 상호 교환 형식을 제공하도록 설계되었습니다. **하둡**에서 제공하는 다른 컬럼형 스토리지 파일 형식(즉, RCFile 및 ORC)과 유사합니다.

Parquet의 특징

- 무료 오픈 소스 파일 형식을 사용합니다.
- 언어를 가리지 않습니다.
- 컬럼 기반 형식 - 파일이 행이 아니라 열로 구성되어, 스토리지 공간이 절약되고 분석 쿼리 속도가 향상됩니다.
- 분석(OLAP) 사용 사례, 그중에서도 기존의 OLTP 데이터베이스와 함께 사용하는 사용 사례에 사용됩니다.
- 데이터 압축과 해제의 **효율이 매우 높습니다**.
- 복잡한 데이터 유형과 고급 중첩 데이터 구조를 지원합니다.
- <https://www.databricks.com/kr/glossary/what-is-parquet>



Row-Based Storage Layout

String	Int	Date
a	1	2020-01
b	2	2020-02
c	3	2020-03



Column-Based Storage Layout

String	Int	Date
a	1	2020-01
b	2	2020-02
c	3	2020-03



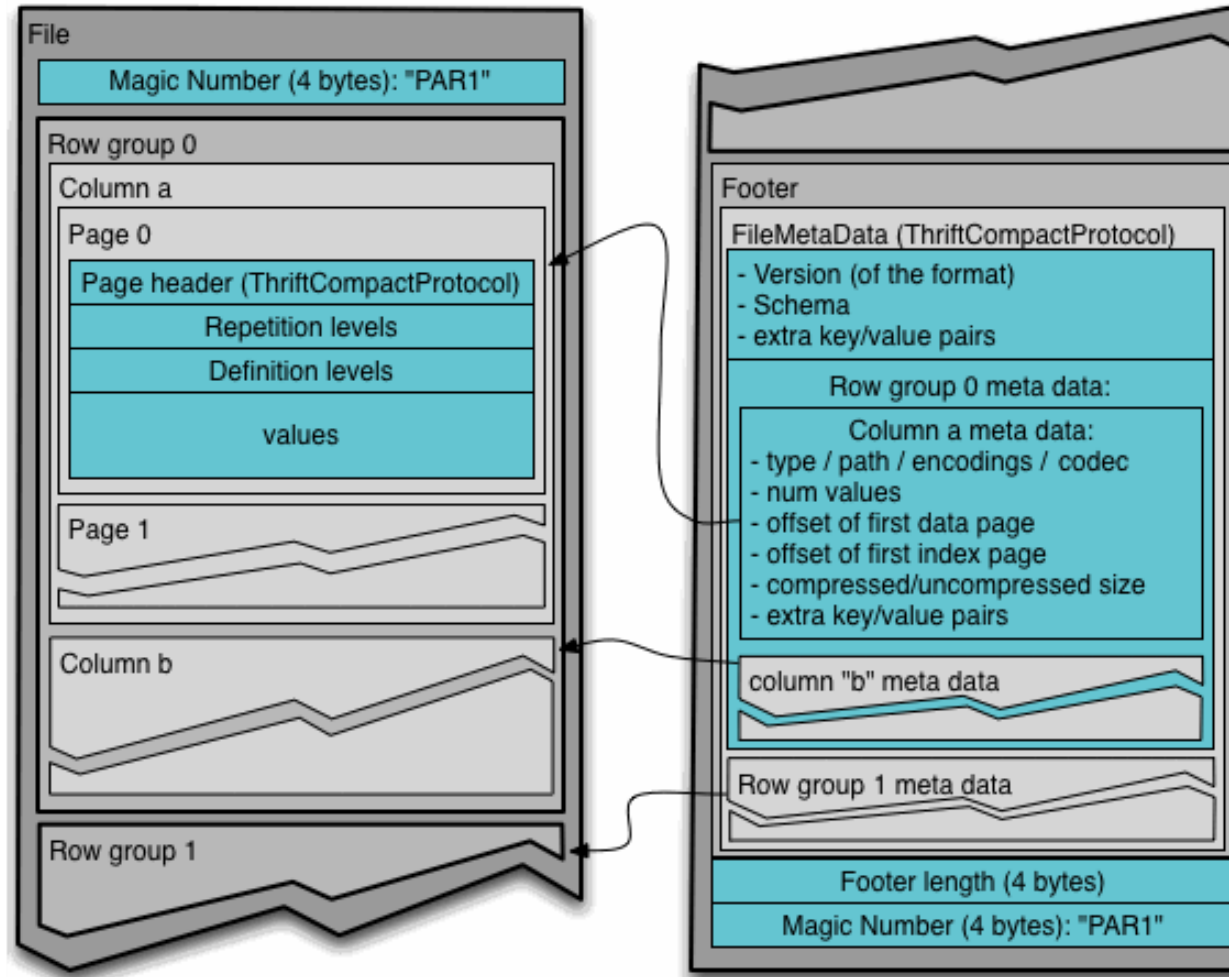
Hybrid-Based Storage Layout (row group size = 2)

String	Int	Date
a	1	2020-01
b	2	2020-02
c	3	2020-03









Final group only has 1 row

- <https://towardsdatascience.com/demystifying-the-parquet-file-format-13adb0206705>



- <https://parquet.apache.org/docs/file-format/>

BIG DATA FORMATS COMPARISON

	Avro	Parquet	ORC
Schema Evolution Support			
Compression			
Splitability			
Most Compatible Platforms	Kafka, Druid	Impala, Arrow Drill, Spark	Hive, Presto
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read

Source: Nexla analysis, April 2018

- <https://towardsdatascience.com/new-in-hadoop-you-should-know-the-various-file-format-in-hadoop-4fcdfa25d42b>

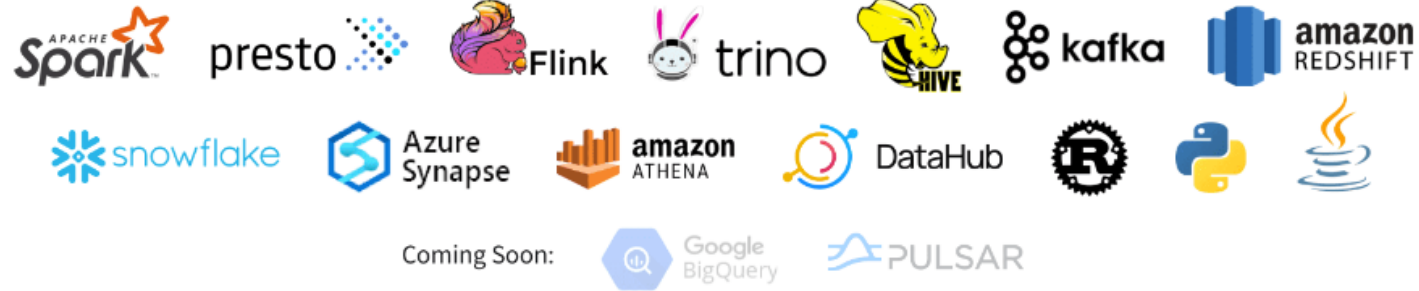
Delta Lake

Delta Lake는 Databricks Lakehouse 플랫폼에 데이터 및 테이블을 저장하기 위한 기반을 제공하는 최적화된 스토리지 계층입니다. Delta Lake는 ACID 트랜잭션 및 확장 가능한 메타데이터 처리를 위해 파일 기반 트랜잭션 로그를 사용하여 Parquet 데이터 파일을 확장하는 오픈 소스 소프트웨어입니다. Delta Lake는 Apache Spark API와 완벽하게 호환되며 구조적 스트리밍과 긴밀하게 통합되도록 개발되어 일괄 처리 및 스트리밍 작업 모두에 단일 데이터 복사본을 쉽게 사용하고 대규모로 증분 처리를 제공할 수 있습니다.

Delta Lake는 Azure Databricks의 모든 작업에 대한 기본 스토리지 형식입니다. 달리 지정하지 않는 한 Azure Databricks의 모든 테이블은 델타 테이블입니다. Databricks는 원래 Delta Lake 프로토콜을 개발했으며 오픈 소스 프로젝트에 지속적으로 기여하고 있습니다. Databricks

- <https://learn.microsoft.com/ko-kr/azure/databricks/delta/>
- <https://delta.io/>

Integrations



Streaming



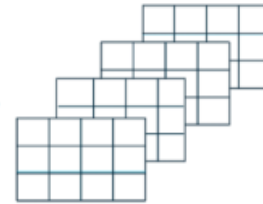
Batch



Ingestion Tables
(Bronze)



Refined Tables
(Silver)



Feature/Agg Data Store
(Gold)

Analytics
and Machine
Learning



Your Existing Data Lake



Azure
Data Lake Storage



amazon
S3



IBM Cloud





Key Features



ACID Transactions

Protect your data with serializability, the strongest level of isolation



Scalable Metadata

Handle petabyte-scale tables with billions of partitions and files with ease



Time Travel

Access/revert to earlier versions of data for audits, rollbacks, or reproduce



Open Source

Community driven, open standards, open protocol, open discussions



Unified Batch/Streaming

Exactly once semantics ingestion to backfill to interactive queries



Schema Evolution / Enforcement

Prevent bad data from causing data corruption



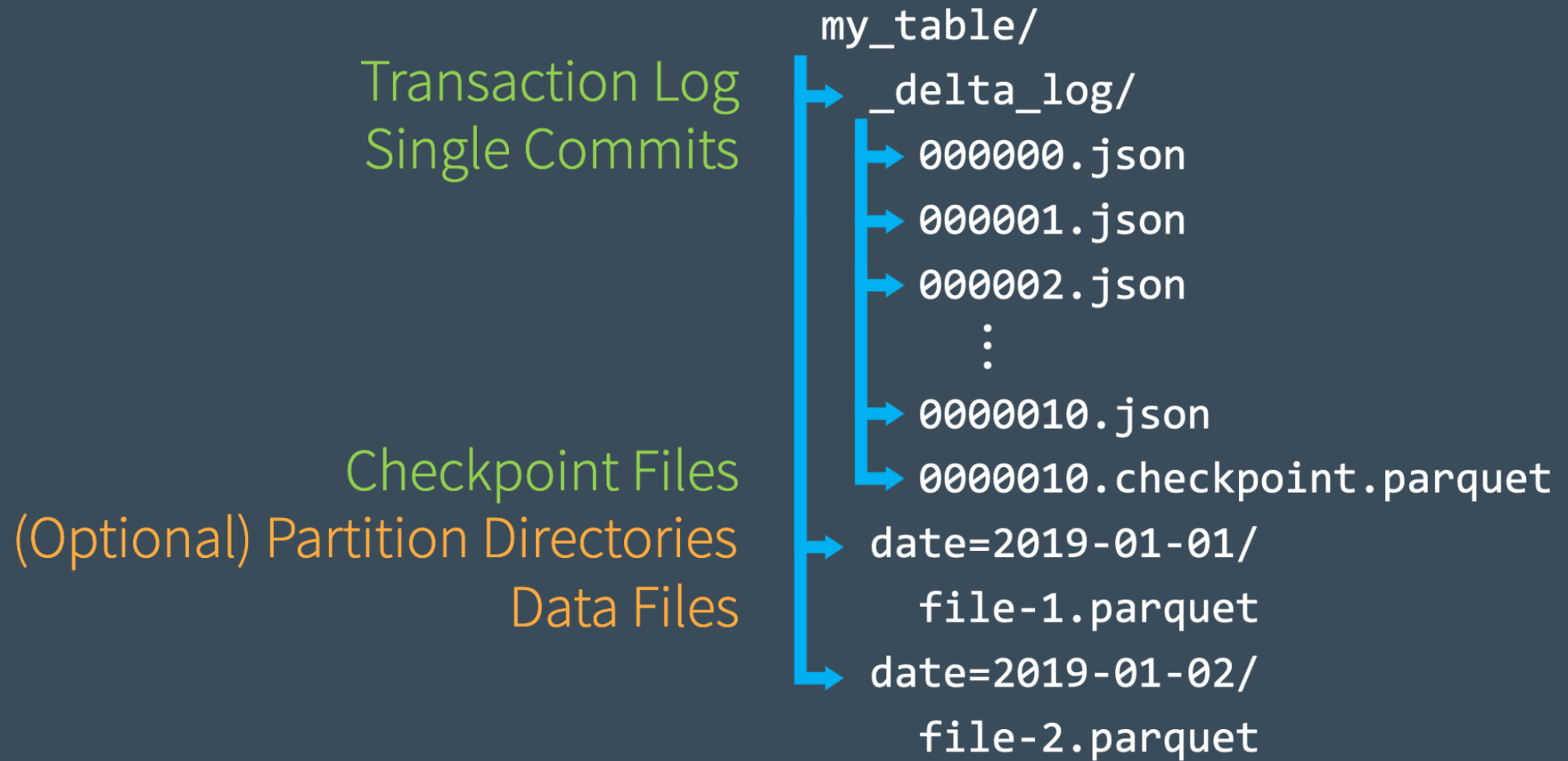
Audit History

Delta Lake log all change details providing a full audit trail



DML Operations

SQL, Scala/Java and Python APIs to merge, update and delete datasets



- <https://www.databricks.com/blog/2019/08/21/diving-into-delta-lake-unpacking-the-transaction-log.html>

Unity 카탈로그

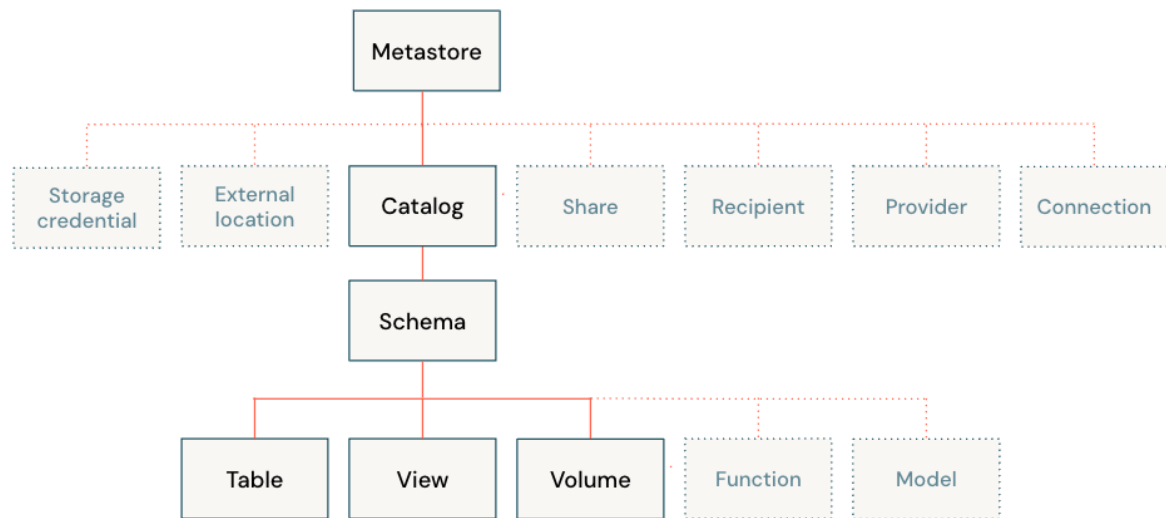
Unity 카탈로그의 주요 기능은 다음과 같습니다.

- **한 번 정의, 모든 곳에서 보안:** Unity 카탈로그는 모든 작업 영역에 적용되는 데이터 액세스 정책을 관리할 수 있는 단일 위치를 제공합니다.
- **표준 규격 보안 모델:** Unity 카탈로그의 보안 모델은 표준 ANSI SQL을 기반으로 하며 관리자는 카탈로그, 데이터베이스(스키마라고도 함), 테이블 및 뷰 수준에서 친숙한 구문을 사용하여 기존 데이터 레이크에서 권한을 부여할 수 있습니다.
- **기본 제공 감사 및 계보:** Unity 카탈로그는 데이터에 대한 액세스를 기록하는 사용자 수준 감사 로그를 자동으로 캡처합니다. 또한 Unity 카탈로그는 모든 언어에서 데이터 자산을 만들고 사용하는 방법을 추적하는 계보 데이터를 캡처합니다.
- **데이터 검색:** Unity 카탈로그를 사용하면 데이터 자산에 태그를 지정하고 문서화할 수 있으며 데이터 소비자가 데이터를 찾을 수 있도록 도와주는 검색 인터페이스를 제공합니다.
- **시스템 테이블(공개 미리 보기):** Unity 카탈로그를 사용하면 감사 로그, 청구 가능 사용량 및 계보를 포함하여 계정의 운영 데이터에 쉽게 액세스하고 쿼리할 수 있습니다.
- <https://learn.microsoft.com/ko-kr/azure/databricks/data-governance/unity-catalog/>

Unity 카탈로그 개체 모델

Unity 카탈로그에서 기본 데이터 개체의 계층 구조는 메타스토어에서 테이블 또는 볼륨으로 흐릅니다.

- **Metastore**: 메타데이터에 대한 최상위 컨테이너입니다. 각 메타스토어는 데이터를 구성하는 3단계 네임스페이스(`catalog.schema.table`)를 노출합니다.
- **카탈로그**: 데이터 자산을 구성하는 데 사용되는 개체 계층의 첫 번째 계층입니다.
- **스키마**: 데이터베이스라고도 하는 스키마는 개체 계층의 두 번째 계층이며 테이블과 뷰를 포함합니다.
- **볼륨**: 볼륨은 개체 계층 구조의 가장 낮은 수준에서 테이블 및 뷰와 나란히 배치되며 테이블 형식이 아닌 데이터에 대한 거버넌스를 제공합니다.
- **테이블**: 개체 계층 구조의 가장 낮은 수준에는 테이블과 뷰가 있습니다.



관리되는 테이블

관리되는 테이블은 Unity 카탈로그에서 테이블을 만드는 기본 방법입니다. Unity 카탈로그는 이러한 테이블의 수명 주기 및 파일 레이아웃을 관리합니다. Azure Databricks 외부의 도구를 사용하여 이러한 테이블의 파일을 직접 조작해서는 안 됩니다.

기본적으로 관리 테이블은 메타스토어를 만들 때 구성하는 루트 스토리지 위치에 저장됩니다. 필요에 따라 카탈로그 또는 스키마 수준에서 관리되는 테이블 스토리지 위치를 지정하고 루트 스토리지 위치를 재정의할 수 있습니다. 관리되는 테이블은 항상 **Delta** 테이블 형식을 사용합니다.

관리되는 테이블이 삭제되면 기본 데이터는 30일 이내에 클라우드 테넌트에서 삭제됩니다.

외부 테이블

외부 테이블은 Unity 카탈로그에서 데이터 수명 주기 및 파일 레이아웃을 관리하지 않는 테이블입니다. 외부 테이블을 사용하여 Unity 카탈로그에 많은 양의 기존 데이터를 등록하거나 Azure Databricks 클러스터 또는 Databricks SQL 웨어하우스 외부의 도구를 사용하여 데이터에 직접 액세스해야 하는 경우.

외부 테이블을 삭제하면 Unity 카탈로그는 기본 데이터를 삭제하지 않습니다. 외부 테이블에 대한 권한을 관리하고 관리 테이블과 동일한 방식으로 쿼리에서 사용할 수 있습니다.

외부 테이블은 다음 파일 형식을 사용할 수 있습니다.

- 델타
- CSV
- JSON
- Avro
- 쪽모이 세공 마루
- ORC
- 텍스트

어떤 프로그래밍 언어?

- 기승전 SQL
- 파이썬
- 자바
- ...

파이썬으로 데이터 분석

- Numpy
- Pandas
- ...

NumPy 

 pandas

Jupyter Notebook이란 무엇입니까?



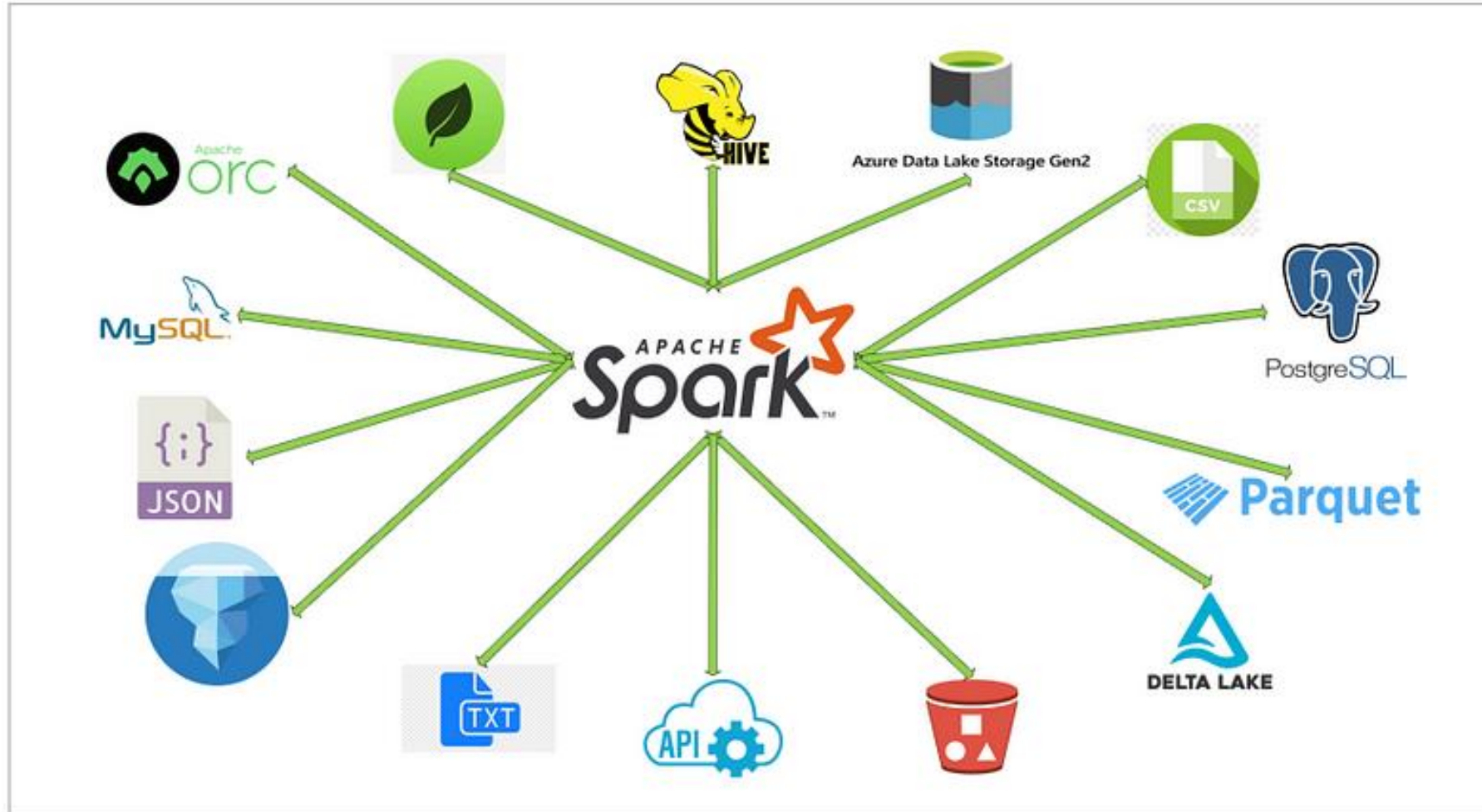
Jupyter Notebook은 오픈 소스 웹 애플리케이션으로, 데이터 사이언티스트가 라이브 코드, 식, 기타 멀티미디어 리소스를 포함하여 문서를 생성 및 공유하는 데 사용할 수 있습니다.

Jupyter Notebook은 어떤 용도로 사용하나요?

Jupyter Notebook은 탐색적 데이터 분석(EDA), 데이터 정리 및 변환, 데이터 시각화, 통계적 모델링, 머신 러닝, 딥 러닝 등의 각종 데이터 사이언스 작업에 사용됩니다.

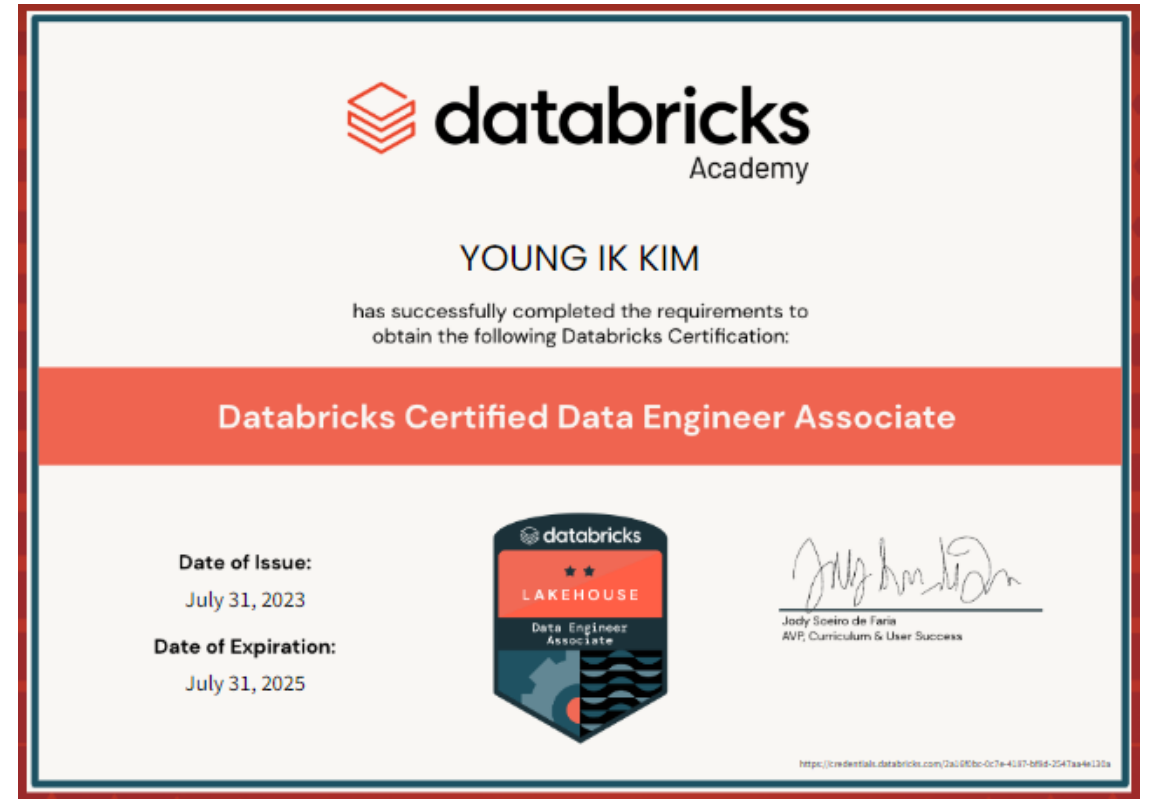
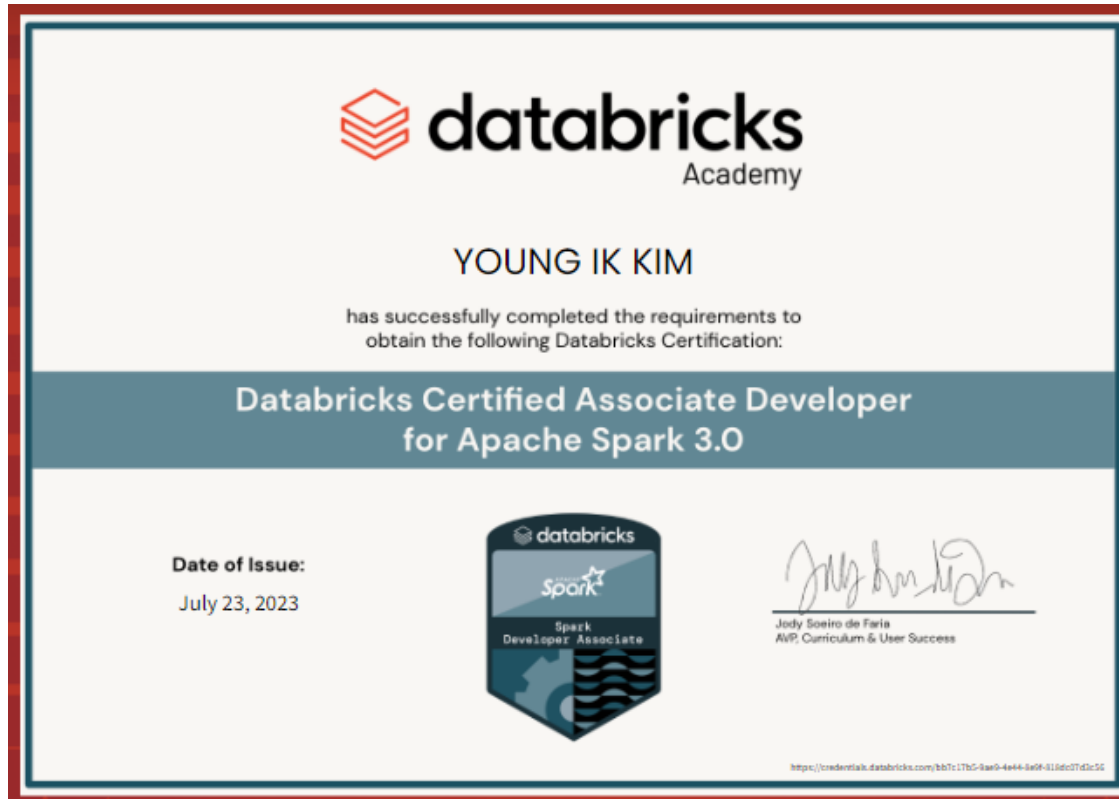
- <https://www.databricks.com/kr/glossary/jupyter-notebook>
- <https://www.anaconda.com/distribution/>





- <https://levelup.gitconnected.com/spark-etl-chapter-9-with-lakehouse-apache-iceberg-38e8fbf20e1>

Databricks 자격증 두개 취득



- <https://www.databricks.com/learn/certification>

단기간 다수의 자격증 취득 역효과

- 스트레스, 예민함



2023. 7. 21 주문

배송완료 • 7/22(토) 도착



🚀로켓프레시 호두마루 (냉동), 70ml, 40입
22,120 원 · 1 개



🚀로켓프레시 체리마루 (냉동), 70ml, 40입
20,000 원 · 1 개

2023. 8. 3 주문

배송완료 • 8/4(금) 도착



🚀로켓프레시 요맘때 바 플레인 (냉동), 70ml, 40개
21,840 원 · 1 개

장바구니

감사합니다