Data Lakehouse Eğitimi

Süre: **3 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Overview

Data Lakehouse Eğitimi, modern veri platformlarında büyük hacimli verilerin ekonomik, güvenilir ve ölçeklenebilir biçimde saklanması, işlenmesi ve analiz edilmesine yönelik mimari yaklaşımları ele alır.

Data Lakehouse mimarisi, geleneksel **Data Warehouse** sistemlerinin analitik yeteneklerini **Data Lake** yapılarının esnekliğiyle birleştirir. Böylece farklı veri kaynaklarından gelen bilgilerin merkezi bir yapıda yönetilmesi, farklı sorgu motorlarıyla analiz edilmesi ve veri işleme süreçlerinin daha etkin yürütülmesi mümkün hale gelir.

Eğitim kapsamında **S3 uyumlu Object Storage**, **Apache Iceberg**, **Medallion Architecture**, **Apache Spark**, **Trino**, **Dremio**, **Kafka**, **CDC**, **OpenMetadata**, **Apache Ranger**, **dbt** ve **Apache Airflow** teknolojileri incelenir.

Yaklaşık **%60 uygulama ağırlıklı** olarak tasarlanan programda katılımcılar; veri yükleme ve dönüştürme süreçleri geliştirme, **batch ve streaming pipeline'ları** oluşturma, sorgu performansını iyileştirme, **metadata ve erişim politikalarını** yönetme ve **CI/CD destekli Lakehouse veri akışları** tasarlama konularında deneyim kazanır.

Prerequisites

Katılımcıların **temel SQL ve Python** bilgisine sahip olmaları, ayrıca **Docker** ve **Kubernetes** teknolojilerine aşina olmaları önerilir.

Who Should Attend

- **Modern veri platformları** geliştirmek isteyen profesyoneller
- **Platform ekipleri**
- **Veri analistleri**
- **DevOps mühendisleri**
- **Kıdemli veri mühendisleri**
- **Veri altyapısı yöneticileri**

Data Lakehouse Eğitimi

Süre: **3 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

What You Will Learn

Eğitim sonunda:

- **Data Warehouse, Data Lake ve Lakehouse** mimarilerini karşılaştırabileceksiniz.
- **S3 uyumlu Object Storage** sistemleri üzerinde veri depolama yapıları tasarlayabileceksiniz.
- **Apache Iceberg** ile ACID, schema evolution ve time travel özelliklerini kullanabileceksiniz.
- **Bronze, Silver ve Gold** katmanlarından oluşan **Medallion Architecture** tasarlayabileceksiniz.
- **Trino ve Dremio** ile analitik ve federated SQL sorguları çalıştırabileceksiniz.
- **PySpark** kullanarak veri pipeline'ları geliştirebilecek ve performans sorunlarını inceleyebileceksiniz.
- **Kafka ve Spark Structured Streaming** ile gerçek zamanlı veri akışları oluşturabileceksiniz.
- **CDC ve MERGE** işlemleriyle değişen verileri Lakehouse tablolarına aktarabileceksiniz.
- **OpenMetadata ve Apache Ranger** ile veri yönetişimi, erişim ve denetim süreçlerini yönetebileceksiniz.
- **dbt** ile test edilebilir veri dönüşümleri ve**Fact, Dimension, Star Schema, SCD** yapılarına uygun veri modelleri geliştirebileceksiniz.
- **CI/CD destekli, üretim ortamına uygun Lakehouse pipeline'ları** tasarlayabileceksiniz.

Outline

1. Lakehouse Mimarisi ve SQL Engine

Data Lakehouse Mimarisi

- **Lakehouse** kavramı ve mimarinin gelişimi
- **Data Warehouse → Data Lake → Lakehouse** dönüşümünün nedenleri
- Düşük maliyetli **Object Storage** üzerinde **ACID** özellikleriyle güvenilir analitik işlemler gerçekleştirme
- **Compute ve storage ayrımı**
- Aynı verilerin **Spark, Trino ve Dremio** gibi farklı motorlar tarafından kullanılabilmesi

S3 Object Storage Yapıları

- **MinIO, AWS S3 ve Ceph** gibi S3 uyumlu depolama çözümleri

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- **On-premises** veri depolama senaryoları
- **Parquet, ORC ve Avro** dosya formatları
- Kolonsal veri depolamanın avantajları
- **Small file problem** ve **partitioning** stratejileri

Açık Tablo Formatları (Open Table Formats)

- Parquet kullanımının tek başına yeterli olmadığı durumlar
- **Schema evolution, ACID ve time travel** gereksinimleri
- **Apache Iceberg, Delta Lake ve Apache Hudi** karşılaştırması
- **Apache Iceberg** mimarisinde snapshot, manifest ve metadata yapıları
- **Hidden partitioning** ve **partition evolution**

Medallion Architecture

- **Bronze:** Kaynak sistemlerden alınan ham verilerin saklanması
- **Silver:** Temizlenmiş, doğrulanmış ve tekrarlardan arındırılmış verilerin hazırlanması
- **Gold:** Raporlama ve iş analitiğine uygun veri katmanlarının oluşturulması
- Katmanlar arasındaki görevlerin ayrılması ve **veri kalitesi kontrolleri**

Apache Spark'a Genel Bakış

- **Spark** teknolojisinin Lakehouse mimarisindeki rolü
- Veri yükleme (**ingestion**) ve dönüştürme işlemleri
- Spark'ın diğer Lakehouse bileşenleriyle ilişkisi

Uygulama 1: Mini Lakehouse Kurulumu

- **MinIO, Apache Iceberg ve Spark** ile temel Lakehouse ortamının hazırlanması
- **Bronze** katmanına ilk verilerin yüklenmesi
- **Time travel** özelliğinin denenmesi

2. SQL Engine ve Analitik Sorgulama

SQL Engine Kavramı

- Bağımsız sorgu motorlarının çalışma prensipleri
- Veriyi taşımadan bulunduğu ortamda sorgulama

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- **MPP (Massively Parallel Processing)** mimarisi
- **Coordinator ve worker** yapıları
- **Federated query** ile farklı veri kaynaklarının tek SQL sorgusunda birleştirilmesi

Trino

- **Coordinator, worker ve connector** mimarisi
- **Iceberg, Hive, PostgreSQL ve Kafka** bağlantılarının yapılandırılması
- Bellek yönetimi
- **Fault-tolerant execution** yaklaşımı
- **Ad-hoc analitik**, iş zekası ve federated sorgulama senaryoları

Dremio

- **Apache Arrow** tabanlı sorgu yürütme mimarisi
- **Columnar Cloud Cache (C3)**
- **Reflection** ile sorgu performansının artırılması
- **Virtual dataset** ve self-service analitik yaklaşımı
- **Apache Polaris** tabanlı entegre katalog ve Iceberg üzerinde yazma desteği

Trino ve Dremio Karşılaştırması

- **Trino'nun sorgu motoru** ve **Dremio'nun entegre platform** yaklaşımı
- Manuel optimizasyon ile **Reflection** tabanlı hızlandırma yöntemleri
- **Veri yönetişimi** ve semantik katman farklılıkları
- Lisanslama, operasyonel gereksinimler ve ekip yetkinliğine göre platform seçimi

Uygulama 2: Iceberg Tablolarında Sorgu Performansı

- Aynı **Apache Iceberg** tablosunun **Trino ve Dremio** üzerinden sorgulanması
- **EXPLAIN** ile sorgu planlarının karşılaştırılması
- Dremio üzerinde **Reflection** oluşturulması
- Sorgu performansındaki değişimin gözlemlenmesi

3. Spark ile Yazılım Geliştirme, Metadata ve Governance

Bu bölümde **Spark tabanlı veri işleme**, **metadata yönetimi** ve **veri yönetişimi** konuları ele alınır.

Data Lakehouse Eğitimi

Süre: **3 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Eğitim kazanımları kapsamında aşağıdaki çalışmalar öne çıkar:

- **PySpark** ile veri işleme pipeline'larının geliştirilmesi
- Veri işleme performansının analiz edilmesi
- **Kafka, Spark Structured Streaming ve Iceberg** ile gerçek zamanlı veri akışları
- **CDC (Change Data Capture)** verilerinin **MERGE** işlemleriyle Lakehouse tablolarına aktarılması
- **OpenMetadata** ile veri keşfi, lineage, profiling ve veri kalitesi yönetimi
- **Apache Ranger** ile erişim kontrolü, kolon maskeleme ve audit politikaları

4. dbt, CI/CD, Airflow ve Veri Modelleme

Bu bölüm, Lakehouse ortamında veri dönüşümlerinin, veri modellerinin ve otomasyon süreçlerinin yönetimine odaklanır.

Eğitim kazanımlarına göre ele alınan temel konular:

- **dbt** ile test edilebilir **Silver–Gold** veri dönüşümleri
- **Fact** ve **Dimension** tabloları
- **Star Schema** ve **Slowly Changing Dimensions (SCD)** yapıları
- **Apache Airflow** ve **CI/CD** kapsamında veri pipeline'larının yönetimi
- Üretim ortamına uygun **uçtan uca Lakehouse pipeline** tasarımı