Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Overview

Bu Eğitim Size Ne Kazandırır?

- **Databricks, Apache Spark ve Delta Lake** üzerinde gelişmiş veri mühendisliği süreçlerini tasarlama ve iyileştirme becerilerinizi geliştirebilirsiniz.
- **Spark Structured Streaming** kullanarak gerçek zamanlı veri işleme, akış toplulaştırma ve veri kalitesi yönetimi konularında deneyim kazanabilirsiniz.
- **Unity Catalog, veri anonimleştirme ve Change Data Feed (CDF)** özellikleriyle veri gizliliği ve değişiklik yönetimi yaklaşımlarını öğrenebilirsiniz.
- **Spark UI, Data Skipping, Liquid Clustering ve Shuffle** gibi konuları inceleyerek veri işleme performansını analiz edebilirsiniz.
- **Databricks Asset Bundles (DABs)** kullanarak projeleri farklı ortamlara dağıtmayı ve **CI/CD süreçlerini** otomatikleştirmeyi öğrenebilirsiniz.
- **Visual Studio Code, Databricks CLI ve GitHub Actions** gibi araçların veri mühendisliği iş akışlarındaki kullanımını keşfedebilirsiniz.

Advanced Data Engineering with Databricks Eğitimi, **Databricks platformunda ileri düzey veri mühendisliği** becerilerini geliştirmek isteyen profesyoneller için hazırlanmıştır.

Program, her biri **4 saat süren 4 modülden** oluşur ve toplam **16 saatlik** bir eğitim içeriği sunar. Eğitim, **veri akışı işleme, veri gizliliği, performans optimizasyonu ve otomatik dağıtım** olmak üzere dört temel alana odaklanır.

İlk modülde **Apache Spark Structured Streaming** ve **Delta Lake** kullanılarak akış verilerinin nasıl işlendiği incelenir. Katılımcılar, **akış sorguları, zaman pencereleri, Watermark mekanizmaları, toplulaştırma işlemleri ve Streaming ETL** konularını öğrenir. Ayrıca veri alımı ve veri kalitesi uygulamaları üzerinde çalışır.

İkinci modül, **Databricks ortamında veri gizliliği ve güvenliği** konularını ele alır. **Unity Catalog**, kişisel verilerin korunması, veri izolasyonu, takma adlandırma, anonimleştirme ve **Change Data Feed (CDF)** ile değişikliklerin aktarılması üzerinde durulur.

Üçüncü modülde **Spark ve Delta Lake performans optimizasyonu** incelenir. Katılımcılar, **Spark UI** aracılığıyla iş yüklerini değerlendirmeyi, performans sorunlarını belirlemeyi ve **Data Skipping, Liquid Clustering, Shuffle, Spill ve Serialization** gibi kavramların uygulama performansına etkilerini anlamayı öğrenir.

Son modülde **DevOps, DataOps ve CI/CD** ilkeleri doğrultusunda Databricks projelerinin otomatik dağıtımı ele alınır. **Databricks Asset Bundles (DABs)** ile proje yapılandırma, doğrulama, farklı ortamlara dağıtım ve çalıştırma süreçleri incelenir. **Databricks CLI, Visual Studio Code ve GitHub Actions** gibi araçlarla geliştirme ve dağıtım iş akışları değerlendirilir.

Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Eğitim boyunca gerçekleştirilen **demolar ve uygulamalı laboratuvar çalışmaları**, katılımcıların Databricks üzerinde veri mühendisliği çözümlerini daha verimli, güvenli ve yönetilebilir biçimde geliştirmelerine yardımcı olur.

Prerequisites

Bu eğitim, Databricks ve veri mühendisliği alanında belirli bir deneyime sahip katılımcılar için tasarlanmıştır.

Katılımcıların aşağıdaki bilgi ve becerilere sahip olması beklenir:

- Gelişmiş veri dönüşümleri için **PySpark API'lerini kullanma deneyimi**
- **Python sınıflarını (Python Classes)** geliştirme konusunda bilgi
- **Veri ambarı (Data Warehouse)** veya **veri gölü (Data Lake)** gibi üretim ortamlarında **SQL kullanma deneyimi**
- **Databricks Notebooks** ve **Cluster yapılandırması** konusunda uygulamalı deneyim
- **Delta Lake tablolarının SQL ile oluşturulması ve yönetilmesi** hakkında bilgi

Bu ön koşullar, aşağıdaki eğitimlerin tamamlanması ve ilgili sertifikaların alınmasıyla karşılanabilir:

- **Data Engineering with Databricks (DBDE)**
- **Apache Spark Programming with Databricks (DBASP)**
- **Databricks Certified Data Engineer Associate**
- **Databricks Certified Associate Developer for Apache Spark**

Ön koşullardan birini veya birkaçını karşılamayan katılımcılar için aşağıdaki hazırlık eğitimleri önerilmektedir. Bu eğitimler **herhangi bir sırayla** alınabilir:

- **DBDE – Data Engineering with Databricks**
- **DBASP – Apache Spark Programming with Databricks**

Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Who Should Attend

Eğitim özellikle aşağıdaki profesyonellere yöneliktir:

- **Databricks üzerinde veri işleme süreçlerini optimize etmek ve ölçeklendirmek** isteyen veri mühendisleri
- **Gerçek zamanlı ve toplu veri işleme (Streaming ve Batch Processing)** süreçleriyle çalışan büyük veri uzmanları
- **Delta Lake mimarileri** ve **Structured Streaming** üzerine çalışan bulut mühendisleri
- **Databricks Certified Data Engineer Professional** sınavına hazırlanmak isteyen kişiler

Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

What You Will Learn

Bu eğitimi tamamladığınızda:

- **Spark Structured Streaming** ve **Delta Lake** teknolojilerinin akış verisi işleme süreçlerindeki rollerini açıklayabileceksiniz.
- **Akış sorgularını** ve akış verisi okuma yapılandırmalarını anlayabileceksiniz.
- **Event Time, Time Windows, Aggregation ve Watermark** kavramlarını veri akışı senaryolarında kullanabileceksiniz.
- **Streaming ETL** süreçlerinde veri alımı ve veri kalitesinin korunmasına yönelik yaklaşımları uygulayabileceksiniz.
- **Databricks veri gizliliği** ve mevzuata uyumluluk gereksinimlerini değerlendirebileceksiniz.
- **Unity Catalog** ile veri güvenliği ve erişim yönetimine ilişkin temel uygulamaları tanıyabileceksiniz.
- **Pseudonymization, Anonymization ve PII Security** yaklaşımlarını açıklayabileceksiniz.
- **Change Data Feed (CDF)** ile değişen kayıtların işlenmesini ve değişikliklerin aktarılmasını anlayabileceksiniz.
- **Spark UI** kullanarak iş yükü performansını analiz edebilecek ve sorunları inceleyebileceksiniz.
- **Data Skipping ve Liquid Clustering** tekniklerinin performans optimizasyonundaki rolünü açıklayabileceksiniz.
- **Skew, Shuffle, Spill ve Serialization** kaynaklı performans sorunlarını tanıyabileceksiniz.
- **Cluster ve Instance türü seçimlerinin** iş yükü performansına etkilerini değerlendirebileceksiniz.
- **DevOps, DataOps, CI/CD ve test süreçlerinin** Databricks veri mühendisliği projelerinde nasıl kullanıldığını açıklayabileceksiniz.
- **Databricks Asset Bundles (DABs)** yapısını, temel bileşenlerini ve proje şablonlarını anlayabileceksiniz.
- **Databricks CLI** ile DAB projelerini yapılandırma, doğrulama, dağıtma ve çalıştırma süreçlerini uygulayabileceksiniz.
- **Farklı hedef ortamlara yönelik DAB dağıtımları** oluşturabileceksiniz.
- **Visual Studio Code** ile Databricks projelerini yerel ortamda geliştirme ve test etme yaklaşımlarını tanıyabileceksiniz.
- **GitHub Actions** ile otomatik dağıtım süreçlerinin temel yapısını anlayabileceksiniz.
- **Databricks projelerinde otomatik dağıtım ve performans iyileştirme** çalışmalarına katkıda bulunabileceksiniz.

Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Outline

1. Databricks Streaming and Delta Live Tables

Bu modülde **Spark Structured Streaming** ve **Delta Lake** kullanılarak akış verilerinin işlenmesi, toplulaştırılması ve veri kalitesinin korunması ele alınır.

Konular

- **Streaming Data Concepts:** Akış verisi kavramları
- **Structured Streaming** teknolojisine giriş
- **Demo:** Bir akış sorgusundan veri okuma
- **Delta Lake üzerinden veri akışı (Streaming from Delta Lake)**
- **Laboratuvar:** Streaming Query Lab
- **Toplulaştırma (Aggregation)**, **zaman pencereleri (Time Windows)** ve **Watermarks**
- **Olay zamanı (Event Time)** ve zaman pencereleri üzerinde toplulaştırma
- **Laboratuvar:** Stream Aggregation Lab
- **Demo:** Watermark kullanarak pencereli toplulaştırma
- **Veri alım desenleri (Data Ingestion Patterns)**
- **Demo:** Auto Load to Bronze
- **Demo:** Stream from Multiplex Bronze
- **Veri kalitesi uygulama deseni (Quality Enforcement Pattern)**
- **Demo:** Quality Enforcement
- **Laboratuvar:** Streaming ETL Lab

Bu modül kapsamında katılımcılar, **Structured Streaming hesaplama modellerini**, akış verisi okuma yapılandırmalarını ve veri akışı ortamlarında **veri kalitesinin korunmasına yönelik yöntemleri** inceler.

2. Databricks Data Privacy

Bu modülde Databricks platformunda **veri gizliliği, mevzuata uyumluluk, veri erişim güvenliği ve değişiklik yönetimi** konuları ele alınır.

Konular

- **Regulatory Compliance:** Düzenleyici gerekliliklere uyumluluk
- **Data Privacy:** Veri gizliliği

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

- **Temel kavramlar ve bileşenler**
- **Veri denetimi (Audit Your Data)**
- **Veri izolasyonu (Data Isolation)**
- **Demo:** Unity Catalog ile verilerin güvenliğinin sağlanması
- **Pseudonymization & Anonymization:** Takma adlandırma ve anonimleştirme
- **Özet ve iyi uygulamalar**
- **Demo:** Kişisel olarak tanımlanabilir bilgilerin (PII) güvenliği
- **Değişen verilerin yakalanması (Capturing Changed Data)**
- **Databricks ortamında veri silme**
- **Demo:** Change Data Feed (CDF) kayıtlarının işlenmesi ve değişikliklerin aktarılması
- **Laboratuvar:** Propagating Changes with CDF Lab

Bu modül, Databricks üzerinde veri mühendisliği faaliyetleri yürüten profesyonellerin**veri gizliliği ve güvenliği gereksinimlerini** anlamalarına ve veri değişikliklerini yönetmelerine odaklanır.

3. Databricks Performance Optimization

Bu modülde **Apache Spark ve Delta Lake iş yüklerinin performansını artırmaya** yönelik teknikler incelenir.

Konular

- **DevOps Spark UI** arayüzüne giriş
- **Temel tasarım yaklaşımına (Designing Foundation)** giriş
- **Demo:** File Explosion
- **Data Skipping** ve **Liquid Clustering**
- **Laboratuvar:** Data Skipping and Liquid Clustering
- **Skew:** Veri dağılımındaki dengesizlik
- **Shuffles:** Veri yeniden dağıtım işlemleri
- **Demo:** Shuffle
- **Spill:** İşleme sırasında verilerin diske taşınması
- **Laboratuvar:** Exploding Join
- **Serialization:** Serileştirme

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- **Demo:** User-Defined Functions (UDFs)
- **Fine-Tuning:** Uygun Cluster seçimi
- **En uygun Instance türlerinin belirlenmesi**

Modül kapsamında **Spark UI ile performans analizi**, uygulama hatalarının incelenmesi ve iş yüklerinin iyileştirilmesi üzerinde durulur. Ayrıca **veri yerleşimi, önbellekleme (Caching) ve Photon** gibi performansla ilişkili konulara yönelik genel bir bakış sunulur.

4. Automated Deployment with Databricks Asset Bundles

Bu modülde **Databricks Asset Bundles (DABs)** kullanılarak veri mühendisliği projelerinin yapılandırılması, dağıtılması ve otomatikleştirilmesi incelenir.

Konular

- **DevOps** ilkelerinin gözden geçirilmesi
- **Sürekli Entegrasyon (CI)** ve **Sürekli Dağıtım/Teslimat (CD)** süreçlerinin değerlendirilmesi
- **Demo:** Eğitim ortamının hazırlanması ve kimlik doğrulama
- **Databricks projelerinin dağıtılması**
- **Databricks Asset Bundles (DABs)** teknolojisine giriş
- **Demo:** Basit bir DAB dağıtımı
- **Laboratuvar:** Basit bir DAB dağıtımı
- **DAB yapılandırmalarında değişken kullanımı (Variable Substitutions)**
- **Demo:** Bir DAB'ın birden fazla ortama dağıtılması
- **Laboratuvar:** DAB'ın birden fazla ortama dağıtılması
- **DAB proje şablonlarına** genel bakış
- **Laboratuvar:** Databricks varsayılan DAB şablonunun kullanılması
- **DABs ile CI/CD projelerine** genel bakış
- **Demo:** DABs ile Sürekli Entegrasyon ve Sürekli Dağıtım
- **Laboratuvar:** DABs ile veri mühendisliği iş akışlarına makine öğrenimi eklenmesi
- **Visual Studio Code (VSCode)** ile yerel geliştirme
- **Demo:** VSCode ile Databricks kullanımı
- **Veri mühendisliği için CI/CD iyi uygulamaları**
- **Sonraki adımlar:** GitHub Actions ile otomatik dağıtım

Advanced Data Engineering with Databricks Eğitimi

Süre: **2 Gün**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Modül boyunca **DevOps ve DataOps** yaklaşımlarının Databricks projelerine uygulanması değerlendirilir.

Katılımcılar, **Databricks REST API, SDK ve CLI** gibi dağıtım araçlarını tanır. **Databricks Asset Bundles** yapısının temel bileşenlerini, klasör düzenini ve farklı hedef ortamlar için yapılandırılmasını inceler.

Ayrıca **Databricks CLI** kullanarak DAB yapılandırmalarına değişken ekleme, yapılandırmaları değiştirme, doğrulama, dağıtma ve çalıştırma süreçlerini öğrenir.

Visual Studio Code, Databricks Asset Bundles projelerini yerel ortamda geliştirme, test etme ve dağıtma açısından ele alınır. Modül, **GitHub Actions ile CI/CD dağıtım süreçlerinin otomatikleştirilmesine** giriş yapılarak tamamlanır.