Analista de Datos Cloudera - Virtual

Descrición

O curso céntrase en Apache Pig, Apache Hive e Cloudera Impala, e ten como obxectivo ensinar aos alumnos para aplicar análises de datos tradicionais e obter a habilidade de xestionar as ferramentas de intelixencia de negocio para o Big Data. Cloudera presenta os datos das ferramentas que os profesionais necesitan para acceder, manipular, transformar e analizar conxuntos de datos complexos utilizando SQL e linguaxes de scripting similares.

Apache Hive fai que os datos multi-estruturados sexan accesibles para os analistas, administradores de bases de datos, e outras persoas sen coñecementos de programación Xava. Apache Pig aplica os fundamentos de linguaxes de scripting familiares para o clúster Hadoop. Cloudera Impala permite, en tempo real, a análise interactiva dos datos almacenados en Hadoop a través dunha contorna de SQL nativo

Obxectivos

Ao finalizar a formación, o participante saberá:

· O xeito na que o ecosistema open source de ferramentas Big Data aborda desafíos que non poden abarcar os RDBMSs tradicionais
· Uso de Apache Hive e Apache Impala para proporcionar acceso mediante o uso de SQL aos datos
· A sintaxe e os formatos de datos que utilizan Hive e Impala, incluíndo as funcións e as subconsultas
· Crear, modificar e borrar táboas, vistas e bases de datos; cargar datos; e gardar os resultados en consultas
· Crear e usar particiones e diferentes tipos de arquivos
· Combinar dous ou máis datasets co uso de JOIN ou UNION, segundo sexa conveniente
· Comprensión detallada das funcións analíticas e as funcións de fiestra e uso de ambas
· Almacenar e consultar estruturas de datos complexas ou anidadas
· Procesar e analizar datos semi-estructurados ou non estructurados
· Técnicas para a optimización das consultas en Hive e Impala
· Estender as capacidades de Hive e Impala coa utilización de parámetros, formatos personalizados de arquivos, SerDes e scripts externos
· Determinar se Hive, Impala, un RDBMS ou unha combinación de todos eles é o mellor para unha tarefa determinada

Exame de certificación incluído:
CCA Data Analyst (CCA159)

Dirixido a

Curso dirixido a analistas de datos, especialistas en intelixencia de negocio, desarrolladores, arquitectos de sistemas e administradores de bases de datos. Requírense coñecementos de SQL e estar familiarizado con comandos de Linux. Aínda que non é obrigatorio, recoméndase o manexo dalgunha linguaxe de scripting (Bash scripting, Perl, Python, Ruby). Non son necesarios coñecementos de Hadoop.

É necesario ter a capacidade de ler textos técnicos en inglés.

Perfil do docente

O noso equipo de formación está composto por persoas con máis de 5 anos de experiencia en áreas de alta especialización técnica nos ámbitos de aplicación. Dispoñen das certificacións oficiais do fabricante (neste caso Cloudera) para impartir estes cursos.

DURACIÓN	48 horas
PROGRAMA	Programación 2018/19
MATRÍCULA	Gratuíta
METODOLOXÍA	Virtual
TIPO	CURSO
CERTIFICACIÓN OFICIAL	Sí

EXAME CERTIFICACIÓN	CCA Data Analyst (CCA159)
BENEFICIOS	
HORARIO	De luns a venres de 16:30 a 20:30 horas.
PERIODO INSCRICIÓN	09/05/2019 - 23/05/2019
PROBA DE SELECCIÓN	28/05/2019, 16:00
PERIODO DOCENCIA	20/06/2019 - 05/07/2019
LUGAR DE DOCENCIA	Edificio localizado na r/Airas Nunes s/n, barrio de Conxo, en
Nº PRAZAS	20 (Mínimo 10)

Temario

Introdución

Hadoop Basics
• Por que Hadoop?
• Aspectos xerais de Hadoop
• Almacenamento de datos: HDFS
• Procesamento de datos distribuído: YARN, MapReduce e Spark
• Procesamento e análise de datos: Hive e Impala
• Integración de datos: Sqoop
• Outras ferramentas de datos de Hadoop
• Explicación do escenario con exercicios

Introdución a Hive e Impala
• Que é Hive?
• Que é Impala?
• Por que usar Hive e Impala?
• Esquema e almacenamento de datos
• Comparación entre as bases de datos Hive e as tradicionais
• Casos de uso

Consultas con Hive e Impala
• Táboas e bases de datos
• Sintaxis básica nas consultas Hive e Impala
• Tipos de datos
• Tocar emprego para realizar consultas
• Emprego de Beeline (Shell of Hive)
• Emprego da Concha de Impala

Operadores comúns e funcións integradas
• Operadores
• Funcións escalares
• Funcións de agregación

Almacenamento de datos
• Almacenamento de datos
• Creación de bases de datos e táboas
• Carga de datos
• Alteración de bases de datos e táboas
• Simplificación de consultas con vistas
• Almacenamento dos resultados das consultas

Almacenamento e rendemento de datos
• Táboas de particionamento
• Carga de datos en táboas particionadas

• Cando empregar o particionamento
• Elección do formato de almacenamento
• Xestión de metadatos
• Control de acceso a datos

Traballar con varios conxuntos de datos
• UNIÓN e Xuntanza
• Xestión de valores nulos en xuntanzas
• Unións avanzadas

Funcións analíticas e funcións de fiestras
• Utilización de funcións analíticas comúns
• Outras funcións analíticas
• Ventás deslizantes

Datos complexos
• Datos complexos con Hive
• Datos complexos con Impala

Análise de texto
• Uso de expresións regulares
• Procesamento de texto con SerDes en Hive
• Análise de sentimentos e n-gramos

Optimización de colmea
• Realización das consultas
• Bucketing
• Indexación de datos
• Hive at Spark

Optimización do impala
• Realización de consultas
• Mellorar o rendemento de Impala

Ampliando Hive e Impala
• Personalizar SerDes e formatos de ficheiro en Hive
• Transformación de datos con scripts personalizados en Hive
• Funcións definidas polo usuario
• Consultas parametrizadas

Elección da mellor opción
• Comparación entre as bases de datos MapReduce, Hive, Impala e relacional
• ¿Cal elixir?

Conclusión