



تحليل البيانات الضخمة باستخدام Apache Spark: برمجة PySpark واستعلامات Spark SQL وضبط الأداء

2 - 6 أغسطس 2027

برشلونة - فندق أربع نجوم في حي إيكسامبلا



تحليل البيانات الضخمة باستخدام Apache Spark: برمجة PySpark واستعلامات Spark SQL وضبط الأداء

المرجع: 9748_121 التاريخ: 2 - 6 أغسطس 2027 المكان: برشلونة - فندق أربع نجوم في حي إيكسامبلا الرسوم: SAR 23500

مقدمة:

يصبح تحليل البيانات الضخمة ضرورة عندما تتجاوز أحجام سجلات النقر والمعاملات والحساسات وملفات السجل قدرة جهاز واحد، فتتدخل التقارير ويلجأ المحللون إلى العينات بدل التاريخ الكامل وينتظرون مستخلصات الليل. تبني هذه الدورة من كور كونسبت مهارة عملية في Apache Spark تشمل التخزين الموزع والتقسيم، وبنية Spark، وإطارات البيانات واستعلامات Spark SQL، وتحويلات PySpark، وتنسيقات الملفات العمودية، وضبط الأداء، والمعالجة المتدفقة، وتعلم الآلة على نطاق واسع. وتجري كل جلسة مختبرا على مجموعات بيانات بحجم عدة غيغابايت، ويخرج المشاركون بمسار تحليلات PySpark متكامل وتقرير ضبط الأداء مبني على سؤال عمل واقعي.

أهداف الدورة:

- الحكم على حاجة مجموعة البيانات إلى المعالجة الموزعة بتقدير الحجم والسرعة والتنوع والموثوقية والقيمة مقابل حدود الجهاز الواحد
- شرح كيفية تنفيذ المهمة عبر المشغل الرئيسي والمنفذين ومدير العنقود ومجدول DAG، وقراءة واجهة Spark UI لتتبع المراحل والمهام
- كتابة شيفرة PySpark واستعلامات Spark SQL لتصفية مجموعات البيانات الكبيرة وربطها وتجميعها وتطبيق دوال النوافذ عليها في جداول Delta Lakeg ORCg Parquet
- تشخيص بطء مهام Spark وتطبيق التقسيم والتخزين المؤقت والربط بالبحث والتنفيذ التكيفي للاستعلامات لتقليل إعادة التوزيع وزمن التشغيل
- بناء استعلام متدفق عبر Structured Streaming ومسار تعلم الآلة عبر MLlib لتوسيع التحليل الدفعي إلى الأحداث الواردة والتنبؤ على نطاق واسع
- تسليم مسار تحليلات PySpark متكامل وتقرير ضبط الأداء يمكن للزملاء إعادة تشغيله ومراجعته وجدولته

الفئة المستهدفة:

- محللو البيانات الذين تجاوزت اعمالهم في SQL والجداول الالكترونية قدرة قاعدة بيانات واحدة او حاسوب محمول
- مهندسو البيانات المسؤولون عن استيعاب بيانات السجلات والاحداث والمعاملات وتحويلها
- مطورو ذكاء الاعمال الذين يعدون مجموعات بيانات مجمعة كبيرة لطبقات التقارير
- مهندسو التحليلات المسؤولون عن المهام الدفعية على العناقيد المشتركة او المنصات السحابية
- المحللون التقنيون في الاتصالات والتجزئة والمالية والطاقة والخدمة العامة ممن يتعاملون مع سجلات عالية الحجم

محاور الدورة:

اليوم 1: اسس البيانات الضخمة ومفاهيم الحوسبة الموزعة

- تقدير الخصائص الخمس للبيانات الضخمة: الحجم والسرعة والتنوع والموثوقية والقيمة
- قائمة قرار حدود الجهاز الواحد مقابل التوسع الافقي
- مقارنة كتل Hadoop HDFS والنسخ المتماثل مع حاويات التخزين الكائني
- نموذج MapReduce والانتقال الى المعالجة داخل الذاكرة
- حصر اعباء العمل: احجام البيانات الحالية ومتطلبات زمن الاستجابة ونقاط الالم

اليوم 2: بنية Apache Spark وبيئة PySpark

- المشغل الرئيسي والمنفذون ومديرو العناقيد: الوضع المستقل وYARN وKubernetes
- سلاسة RDD والتقييم المؤجل ومجدول DAG
- تتبع المهام والمراحل والمهام الفرعية والاقسام في واجهة Spark UI
- اعداد SparkSession وبيئة الدفاتر التفاعلية واساسيات spark-submit
- واجهتا Dataset وDataFrame مقارنة بواجهة RDD

اليوم 3: اطرار البيانات واستعلامات Spark SQL وتحويلات PySpark

- قراءة ملفات CSV وJSON وParquet وORC وكتابتها بمخططات صريحة
- التحويلات الضيقة والواسعة: select filterg withColumng joining
- تجميعات groupBy ودوال النوافذ والدوال المعرفة من المستخدم
- العروض المؤقتة في Spark SQL وخطط الاستعلام في محسن Catalyst عبر explain
- جداول Delta Lake: كتابات ACID وفرض المخطط وسجل الاصدارات

اليوم 4: ضبط الاداء والمعالجة المتدفقة وتعلم الالة على نطاق واسع

- تحديد حجم الاقسام coalesce repartitioning واعدادات اقسام اعادة التوزيع
- التخزين المؤقت ومستويات الاستبقاء ومتى يستخدم unpersist
- الربط بالبيث وانحراف البيانات والتنفيذ التكيفي للاستعلامات
- الدفعات المصغرة في Structured Streaming والمصادر والمصبات ونقاط الحفظ
- مسارات MLlib: المحولات والمقدرات وتدريب النماذج الموزع

اليوم 5: مشروع المختبر وتقرير مسار التحليلات

- نظرة عامة على خدمات Spark المدارة وخيارات التخزين الكائني السحابي
- مختبر مسار النقر في التجارة الالكترونية: استيعاب الاحداث الخام وتنقيتها وتقسيمها
- مختبر عدادات المرافق: تقسيم الجلسات والتجميع بالنوافذ الزمنية
- بناء مسار تحليلات PySpark متكامل وتقرير ضبط الاداء
- مراجعة الاقران للشيفرة وعرض المسار

المهارات التي ستكتسبها:

- المعالجة الموزعة للبيانات
- البرمجة باستخدام PySpark
- الاستعلام عبر Spark SQL
- تصميم التخزين العمودي
- ضبط اداء مهام Spark
- معالجة البيانات المتدفقة
- تعلم الالة القابل للتوسع
- التشخيص عبر Spark UI

لماذا تحضر هذه الدورة:

- العودة بمسار تحليلات PySpark متكامل وتقرير ضبط الاداء بني وروجع خلال المختبرات
- تحليل التاريخ الكامل بدل العينات بنقل الاستعلامات الثقيلة من ادوات الجهاز الواحد الى Spark
- خفض زمن تشغيل المهام وتكلفة العنقود باكتشاف مشكلات اعادة التوزيع والانحراف والملفات الصغيرة قبل وصولها الى بيئة الانتاج
- التطبيق على بيانات مسار النقر والعدادات مع محللين ومهندسين من قطاعات متعددة

الخاتمة:

يحول Spark البيانات التي كانت تحتاج الى عينات ومستخلصات ليلية الى بيانات يستعلم عنها المحلل في دقائق، بشرط ان تراعي الشيفرة وتخطيط التخزين طريقة عمل المعالجة الموزعة. تنتقل الدورة من الخصائص الخمس للبيانات الضخمة والتخزين الموزع، الى بنية Spark واطارات البيانات واستعلامات Spark SQL وتحولات PySpark، ثم ضبط الاداء والمعالجة المتدفقة وMLlib. ويجمع اليوم الاخير ذلك في مختبري مسار النقر والعدادات لينتج مسار تحليلات PySpark متكاملًا وتقرير ضبط الاداء لبيئة العمل.