分散式系統 期中報告 Apache Spark

108703005 資科三 林藝潔

1. 問題（目的）與背景

* Apache Spark的核心概念

Apache Spark透過記憶體存取中間運算結果，加強了資料處理的即時性，也達成高容錯性和高伸縮性。Spark還基於DAG（有向無環圖）分配任務，構建平行計算框架。此外，採用RDD（Resilient Distributed Dataset）擔任最小運算單位，得以在讀取等操作時，精確到每筆紀錄。

Spark可以支援Scala、Java和Python的程式撰寫，還能執行在任何Hadoop資料來源上，讓使用者可以輕易轉移已有的資料，塑造出極高的易用及普遍性。

Spark也在核心基礎上，開發出各類型的API。像是能夠進行Stream processing（流處理程式）的Spark Streaming；提供Graph Computation（圖譜計算）服務的GraphX；支援Machine Learning（機器學習）的ML Base和MLlib 庫；協助結構化資料SQL查詢和分析的Spark SQL和Shark。

Apache Spark是個完整的大數據分析生態系統，其便利性和低延遲性使它成為眾多資料科學家、工程師愛用的大數據處理平台。

* Apache Spark的應用場景

在Apache Spark中，使用者只需建立一個系統，就能完整地進行所有資料分析可能遇到的步驟，不用再像過去一樣，使用不同的軟體並花時間及心力做資料格式的轉換。因此，許多知名企業都採用Spark來架構內部的資料分析系統。

像是Uber使用Spark構建了Uber Spark Compute Service (uSCS)，幫助進行Uber乘客和Uber駕駛者的定價運算、需求預測以及餐廳推薦等商業任務，還有協助系統的ETL操作和數據探索，讓uSCS這個管理過億活躍用戶的大數據平台能更加穩定。

還有，騰訊旗下的「廣點通」也是使用Spark的應用之一。騰訊大數據使用了Spark平台來支援採擷分析類別計算、互動式即時查詢以及允許誤差範圍的快速查詢計算。目前有超過200台的Spark叢集，並且累積了大量的案例和營運經驗，已成功應用在廣點通pCTR投放系統上，支援了每日上百億的請求量。

此外，淘寶使用了Spark得以解決複雜的機器學習，像是得多次反覆運算或高計算複雜度的演算法，並將其運用在阿里搜索、廣告業務和淘寶的推薦相關演算法。此外，還使用GraphX解決了許多生產問題，像是採用「最大連通圖」實現社區發現；以「隨機遊走」為基礎分析使用者屬性傳播；使用「度分析」找出中樞節點等計算場景。

Yahoo則是採用Spark叢集，取代原先的商業BI、OLAP工具。Spark承擔報表功能的同時，還能達成更高品質的大數據SQL服務。Yahoo也將此運用在一種讓廣告能尋找到其適合對象的應用：Audience Expansion的演算法中，讓企業能夠更有效率的投放廣告。

Spark不僅是在商業面上，還是系統效能上，都能讓企業或是使用者得到卓越的成效。而且在如今app風行的時代，更讓Spark能夠大展風采，讓大數據得以快速流通，並在高速且極小誤差下處理、分析。