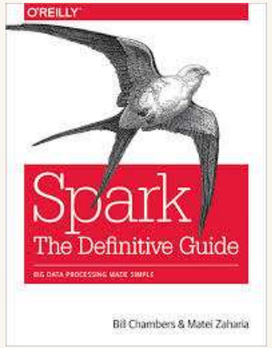
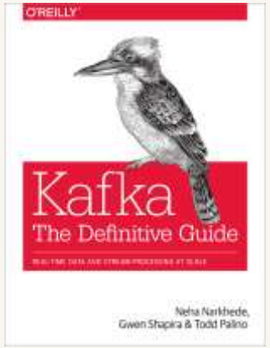
About Me

- 蔡政廷(Eggsy)
- Swipr – Staff Engineer
- Grindr – Senior Data Engineer
- Garmin – Senior Data Engineer
- TripleOne Tech – Technical Leader
- Interestd in SMACK architecture
  - Spark， Cassandra， Kafka， ELK， etc
  - Python， Scala， Java， etc
- 中壢資策會， 台北資策會， 台灣資料工程協會，
  中央大學， 淡江大學講師
- O’ Reilly Translator





www.linkedin.com/in/eggtsai

- 資料倉儲 vs. 資料庫
- Hadoop與Ecosystem

  (批量資料處理及應用實例)
- Apache Spark 簡介

  (串流資料處理及應用實例)
- 大數據產業應用

資料倉儲 vs. 資料庫

(Data Warehouse VS. Database)

資料倉儲和資料庫都是存儲和管理資料的系統，但是它們的設計目的和架構不同。資料庫是為交易而生的，而資料倉儲則是為分析而生的。

如果你需要對數據進行快速讀取和更新，那麼資料庫可能更適合你的需求。而如果你需要對大量數據進行複雜的分析和報告，那麼資料倉儲可能更適合你的需求。

資料倉儲 - 為分析而生

目的

資料倉儲的主要目的是為了支持企業的決策和分析。它會整合來自不同系統的資料，並以主題為導向的方式加以組織和存儲，以便進行跨業務線的分析和報告。

架構

典型的資料倉儲包括數據擷取、轉換和載入(ETL)流程，用於整合各種異質的資料源。此外，它還使用維度建模技術，將資料分類為事實和維度，以支援複雜的分析查詢。

應用

資料倉儲常用於支援業務分析、預測性分析、績效管理和決策支援等需求。它可以幫助企業洞悉客戶行為、優化營運效率和制定更明智的策略決策。

資料庫 - 為交易而生

1 目的

資料庫的主要目的是管理和保護日常業務操作中產生的交易性資料。它必須確保資料的完整性、一致性、可靠性和安全性。

2 架構

資料庫通常使用關係模型，將資料存儲在各種表格中。它提供SQL等語言供開發人員查詢和操作資料，並確保遵守ACID(原子性、一致性、隔離性和持久性)原則。

3 應用

資料庫適用於需要即時、可靠資料處理的應用，如線上交易系統、客戶關係管理(CRM)和會計等。它們確保在高並發和錯誤情境下資料的完整性和一致性。

資料倉儲 vs. 資料庫 - 異同比較

資料模型

資料倉儲採用主題導向的維度模型，以支援複雜分析查詢。相比之下，資料庫則使用關係模型，主要關注交易型業務操作。

資料來源

資料倉儲整合來自多個異質系統的資料，而資料庫通常只包含單一系統產生的交易性資料。

更新頻率

資料倉儲的資料更新通常採用批次處理，以支援複雜的商業智慧分析。相比之下，資料庫需要提供即時更新以確保交易處理的一致性。

資料倉儲與資料庫的未來

1

資料整合

隨著企業數據源的持續增加，資料倉儲和資料庫需要更好地整合，提供統一的數據視圖和分析能力。

2

實時分析

企業需要更快速的洞察和決策反應，資料倉儲需要提供更即時的分析功能，而資料庫也需增強實時處理能力。

3

人工智能

人工智能技術可以增強資料倉儲和資料庫的自動化和智能分析能力，提升企業的數據驅動決策能力。

結論

資料倉儲和資料庫是企業數據管理和分析的兩大基石。雖然兩者有著不同的設計目的和架構特點，但未來它們必將趨於融合，為企業提供更強大、更智能的數據服務。資訊專業人士需要深入了解這兩種技術的特點，並根據業務需求靈活運用，以推動企業向數據驅動型轉型。

Hadoop簡介

Hadoop是一個開源的大數據框架，提供分散式計算和儲存的解決方案。

Hadoop的發展歷程

1

起源

map : 分散出去
reduce : 回收回來

2004年，Google發表了GFS和MapReduce論文。

google file system

2

開源化

2005年，Doug Cutting創建了Hadoop開源專案。

3

成熟期

2010年以後，Hadoop得到了廣泛的企業採用。

Hadoop的核心組件

HDFS

分散式檔案系統，為Hadoop提供可靠的資料存儲。

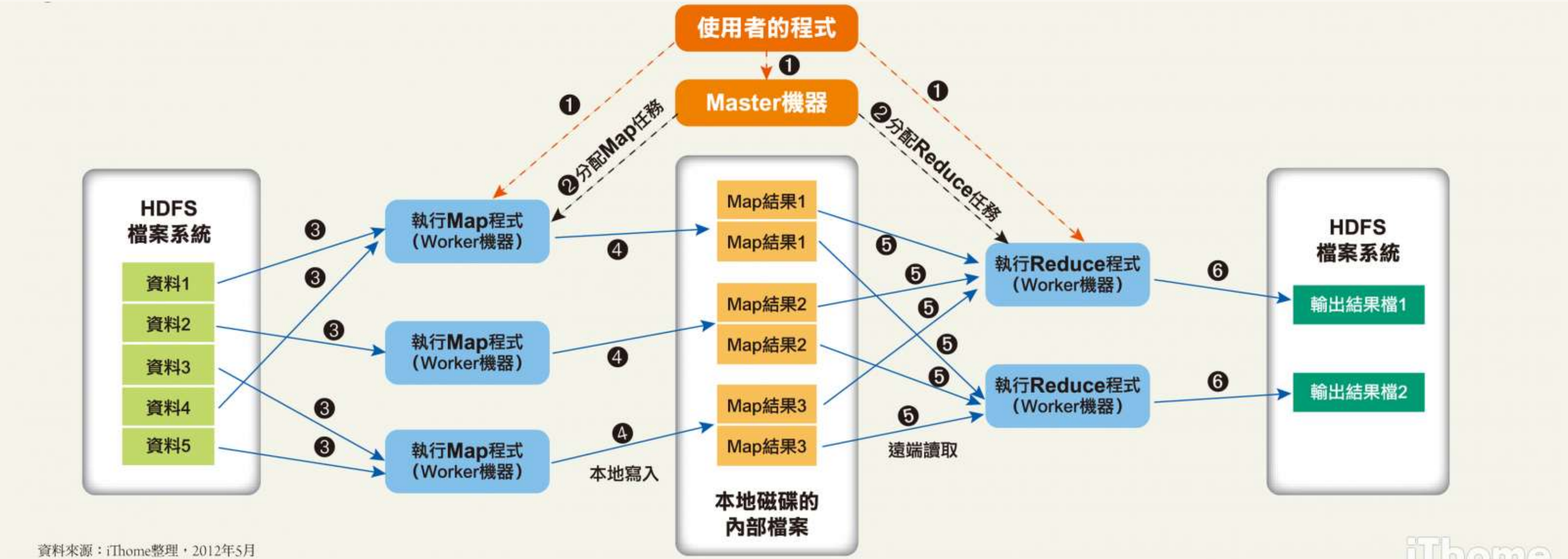
MapReduce

並行計算框架，用於大數據的處理與分析。

YARN

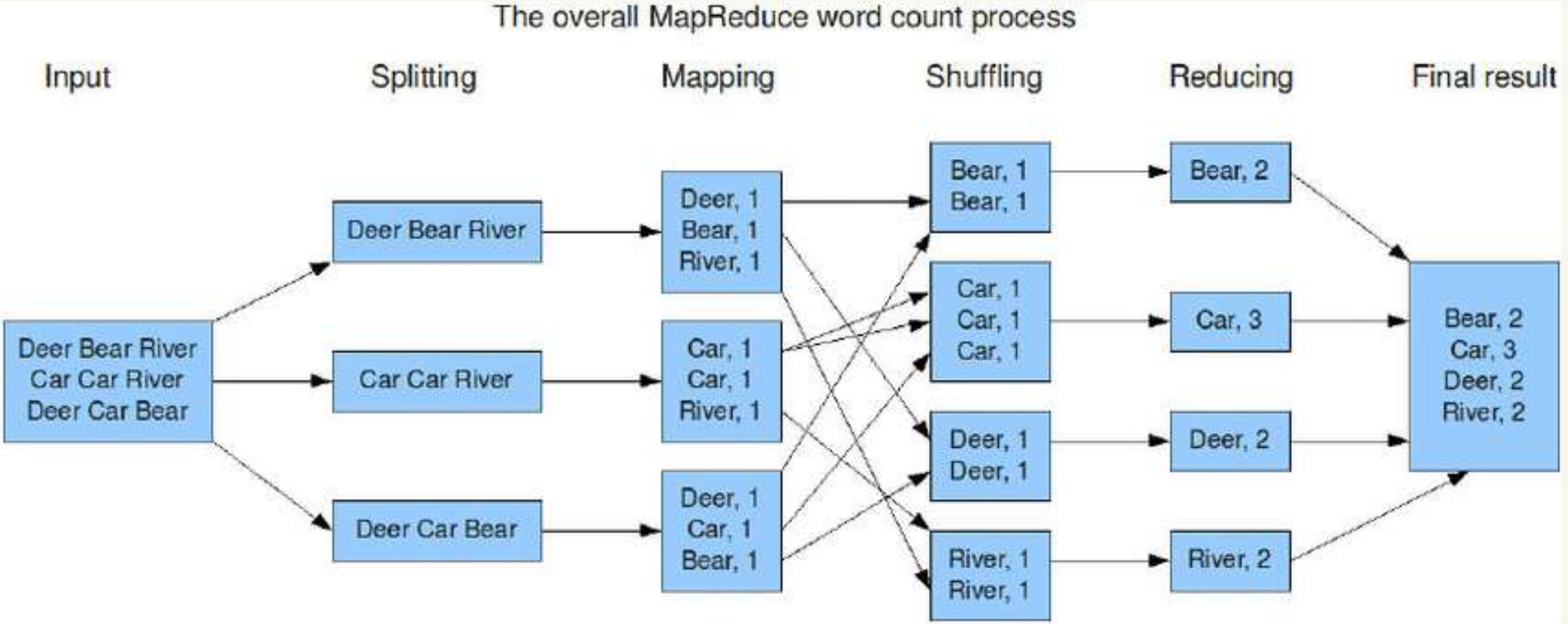
資源管理系統，負責作業調度和集群資源管理。

MapReduce



資料來源：iThome整理，2012年5月

Word Count Example



Hadoop的部屬方式

1

內部部署

在自有硬體上建立Hadoop叢集。

2

私有雲

使用私有雲基礎設施部署Hadoop。

3

公有雲

運用公有雲服務托管Hadoop叢集。

Hadoop的優勢

1

海量資料處理

可靠地處理TB甚至PB級的大量非結構化資料。

2

高可用性

能夠自動容錯和負載平衡，確保系統高可用。

3

橫向擴展

通過增加商用硬體節點來擴展計算能力。

Hadoop生態系統

Spark

快速的大數據分析引擎

Hive

提供SQL語法的資料倉儲工具

Kafka

分散式串流處理平台

HBase

分散式、可擴展的NoSQL資料庫

Hadoop的未來發展

1

大數據分析

Hadoop在海量資料分析方面將持續發揮重要作用。

2

雲端運算

Hadoop可以與雲端服務緊密整合，提供彈性伸縮。

3

物聯網

Hadoop將成為物聯網大數據處理的重要基礎。

結語

Hadoop是一個強大的大數據分析框架，在很多行業都有廣泛應用。隨著Hadoop Ecosystem的不斷發展，它將繼續為企業提供更多創新的大數據解決方案。

批量資料處理及應用實例

在資訊時代，大量資料已成為組織營運和決策的關鍵要素。如何有效地處理和應用這些數據，成為各行各業亟需解決的問題。

批量資料處理的技術要點

1

資料收集

良好的資料收集是批量處理的基礎。這包括確保數據源的可靠性和多樣性，採用自動化採集技術，以及對資料格式和結構進行標準化。

2

資料清洗

在開始處理前，需要對資料進行徹底的清理和整理，消除錯誤、重複和不完整的項目，以確保數據的完整性和準確性。

3

資料分析

利用統計、機器學習等先進技術，對大量資料進行深入的分析和挖掘，找出隱藏的模式和洞見，為後續的決策提供支持。

批量資料處理的應用實例

金融業

金融機構利用批量資料處理技術進行風險評估、投資組合分析、欺詐偵測等，提高決策的準確性和效率。如華爾街投資公司利用機器學習模型實時監測全球交易動向，及時識別並應對金融風險。

零售業

零售企業運用大數據分析顧客行為模式，優化產品推薦、庫存管理、定價策略等，提升營運效率和顧客體驗。如亞馬遜利用人工智慧分析客戶數據，為每位用戶提供個性化的購物建議。

物流業

物流公司採用批量資料處理技術，改善配送路線規劃、貨物追蹤、運輸成本管控等，提高整體運營效率。如UPS利用數據分析和優化演算法，使配送路線縮短6.2%，節省數百萬美元成本。

跨領域創新應用

智慧城市

城市管理部門結合批量資料處理技術，整合交通、能源、環境等多方面資訊，優化城市運作，提高民眾生活品質。如新加坡利用大數據分析改善交通規劃，減少擁堵，提升空氣品質。

醫療保健

醫療機構採用批量資料處理，加強疾病預防、個人化治療方案制定等，提升診斷及治療效果。如IBM Watson利用機器學習技術分析龐大的醫療數據，為癌症患者提供個性化的治療建議。

農業

農業生產商利用大數據分析，優化種植管理、供應鏈物流等，提高生產效率和產品品質。如澳洲農場利用物聯網數據監測作物生長情況，並通過大數據分析，制定精準的灌溉和施肥計劃。

Apache Spark 簡介

Apache Spark是一個快速、通用、可擴展的大數據分析引擎。它能夠在內存中快速運行大量分析任務。

什麼是Spark

Apache Spark是一個開源叢集運算框架，最初是由加州大學柏克萊分校AMPLab所開發。相對於Hadoop的MapReduce會在執行完工作後將中介資料存放到磁碟中，Spark使用了記憶體內運算技術，能在資料尚未寫入硬碟時即在記憶體內分析運算

Hadoop 問題點

- Google 的 MapReduce，他是一個簡單通用與自動容錯的 批次處理計算模型。
  https://static.googleusercontent.com/media/research.google.com/zh-TW//archive/mapreduce-osdi04.pdf

- 由於 Hadoop 很多子專案都繼承 MR 模型，對於其他類型計算，諸如互動式與串流式計算，並不適合被拿來使用。導致大量不同於 MR 的 專有資料處理模型出現，Ex:Storm、Impala與GraphLab。

- 然而隨著新模型的不斷出現，似乎對於巨量資料處理而言，不同類型的作業應該需要一系列不同的處理架構才可以極佳地完成。但是這些專有系統也有一些不足之處。

Spark 巨量資料處理架構

- Spark 採用一個 RDD 概念（一種新的抽象的彈性資料集），在某種程度上 Spark 是對 MapReduce 模型的一種擴充。
- 主要是把 MR 不擅長的計算工作（反覆運算、互動式與串流式）進行改善，並提出一個統一的引擎。
- 因為 MR 缺乏一種特性，就是在平行計算的各個階段劑型有效的資料共用，這種共用就是 RDD 的本質。利用這種本質來達到這些運算模式。
- 在叢集處理的容錯方式，不像Hadoop將計算建置成一個無環圖的工作集

Spark 的特點

高速

Spark在記憶體內運算更快，處理速度可達傳統Hadoop的100倍。

通用

支援批次處理、即時處理、機器學習等多種應用場景。

可擴展

可以在大型集群中運行，處理tb甚至pb級數據。

用Spark有哪些好處？

- Java、Scala、Python 和 RAPIs。
- 可擴展至超過 8000 個結點。
- 能夠在記憶體內緩存資料集以進行互動式資料分析。
- Scala 或 Python 中的互動式命令列介面可降低橫向擴展資料探索的反應時間。
- Spark Streaming 對即時資料串流的處理具有可擴充性、高吞吐量、可容錯性等特點。
- Spark SQL 支援結構化和和關聯式查詢處理（SQL）。
- MLlib 機器學習演算法和 Graphx 圖形處理演算法的高階函式庫。

Spark 的組成

1 Spark Core

提供內存計算、容錯機制、調度等核心功能。

2 Spark SQL

用於結構化數據處理，支援SQL查詢。

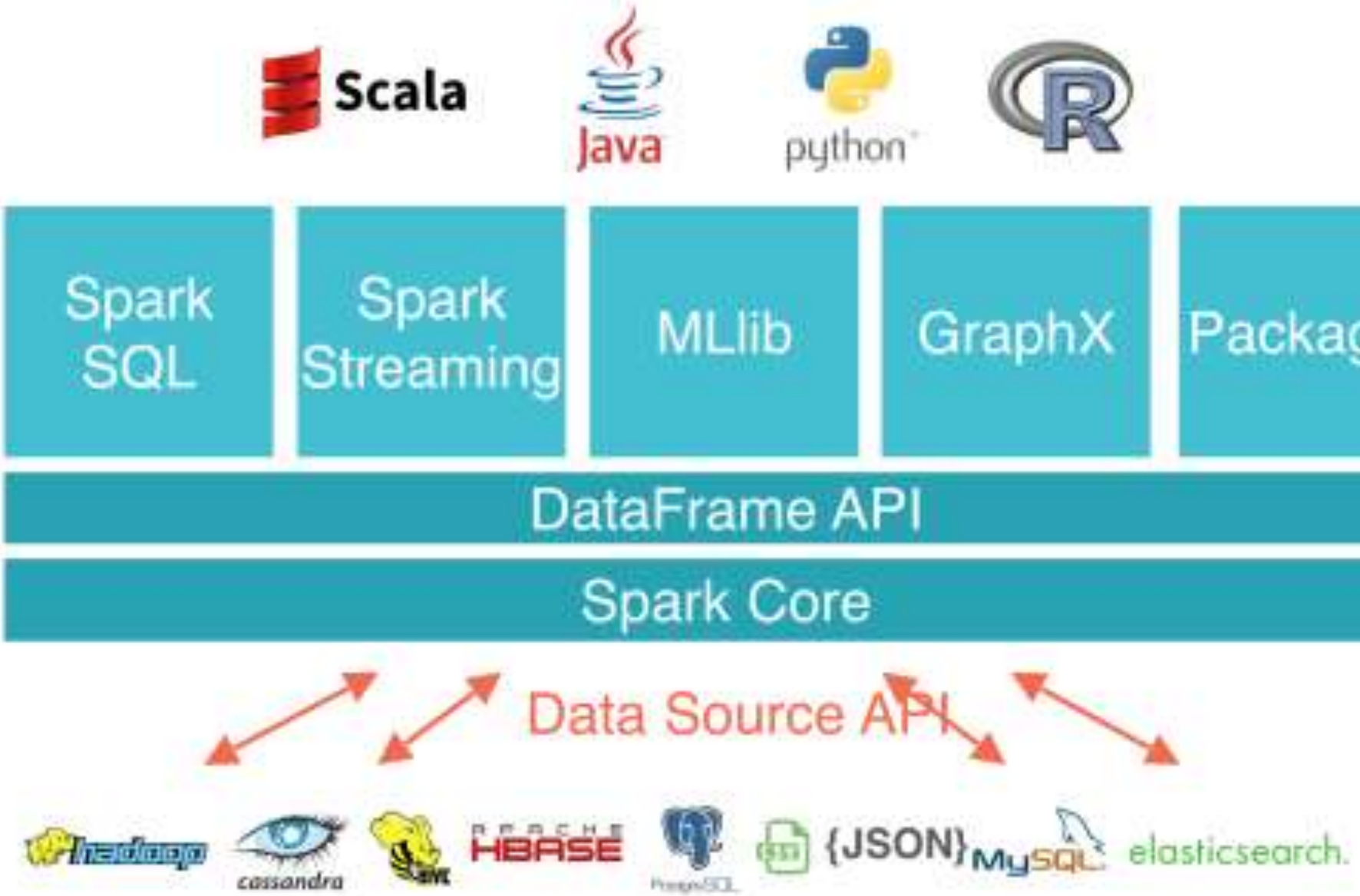
3 Spark Streaming

用於處理實時數據流，支援微批處理。

對應處理的基本情況

巨量資料處理分為以下三種情況：

- **複雜的批次資料處理**（batch data processing）：**時間**長，**跨度為** 10min - N hr。
- **以歷史資料為基礎的互動**查詢（interactive query）：**時間通常為** 10 sec - N min。
- **以即時資料流為基礎的資料處理**（streaming data processing）：**時間 通常為** N ms - N sec。


Scala
Java
python
Spark SQL
Spark Streaming
MLlib
GraphX
Packages
DataFrame API
Spark Core
Data Source API
hadoop
cassandra
HBASE
{JSON}
MySQL
elasticsearch.

Spark部署方式

雲端部署

支援AWS、Azure、GCP等主流雲平台。

自建集群

支援Hadoop YARN、Mesos等集群管理系統。

本地執行

適合開發測試，也可在單機上運行。

Spark 編程模型

1 Resilient Distributed Datasets (RDD)

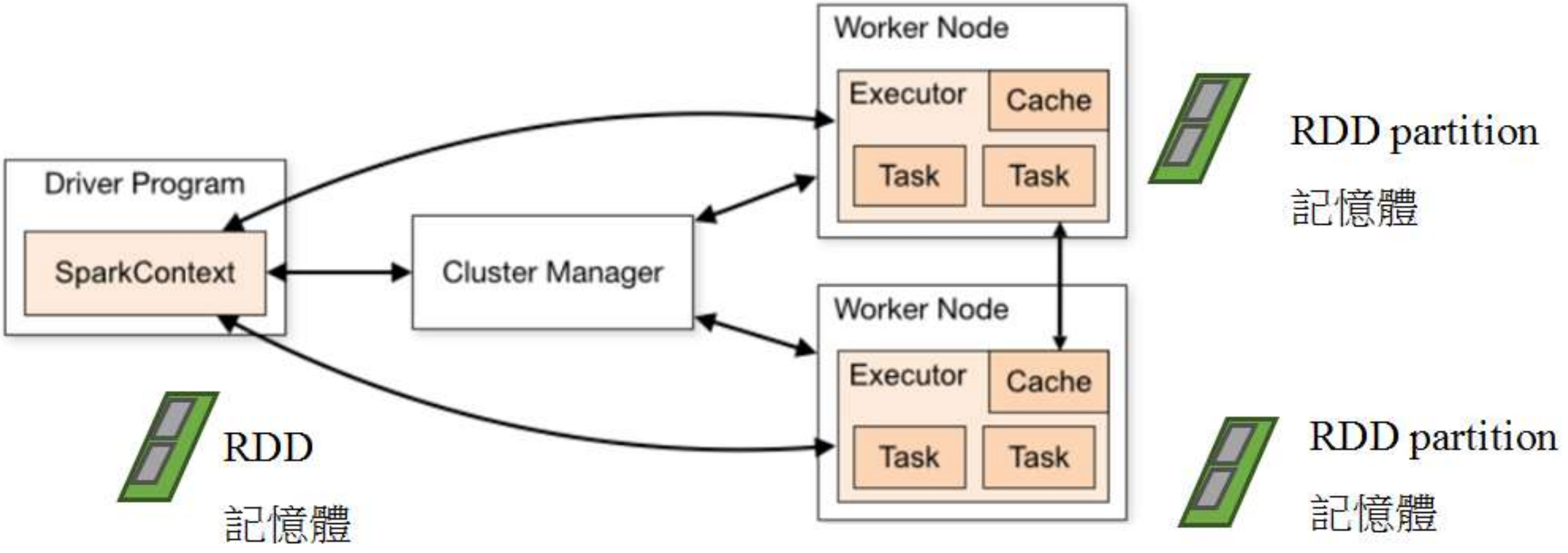
Spark的基礎抽象，彈性分布式數據集。

2 Dataframes

結構化數據的高階抽象，提供SQL語法操作。

3 Datasets

Dataframes的強類型版本，支援面向對象編程。


Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task
RDD partition
記憶體
RDD
記憶體
RDD partition
記憶體

Spark 應用場景

批量數據處理 (SparkSQL)

運行ETL流程、資料探索和分析。

實時數據處理 (Spark Streaming)

分析實時數據流，支持複雜計算。

機器學習和人工智能 (MLLib)

具有豐富的機器學習和深度學習函式庫。

圖計算 (GraphX)

提供高效的圖計算和圖分析功能。

串流資料處理及應用實例

Apache Spark 是一個強大的大數據處理框架，能夠快速高效地處理各種結構化和非結構化的數據，適用於批次處理和串流處理。本文將深入探討 Apache Spark 的串流處理能力，並提供實際應用案例，為您展示如何在企業中運用 Spark 來處理即時資料流，滿足業務需求。

理解 Spark Streaming

1

概念介紹

Spark Streaming 是 Apache Spark 的一個模組，它提供了一種高度可擴展、容錯且易於使用的串流處理引擎。透過將資料流切分為微批次，並以近乎實時的方式進行處理，Spark Streaming 實現了快速、可靠的串流處理功能。

2

核心特性

Spark Streaming 的核心特性包括：1) 可擴展性 - 能夠靈活地處理大規模資料流; 2) 容錯性 - 即使發生故障也能夠自動恢復; 3) 整合性 - 可與 Spark 的其他模組如 Spark SQL 緊密整合。這些特性使 Spark Streaming 成為處理即時資料的理想選擇。

3

應用場景

Spark Streaming 適用於各種即時資料處理場景，如網路流量分析、社交媒體監控、物聯網感測器數據處理、欺詐檢測等。通過高效的串流處理，企業可以即時獲取洞見，做出更快速的決策響應。

Spark Streaming的程式設計模型

1

資料輸入

開發者需要選擇合適的輸入源，如 Kafka 或 Kinesis，並設置相關參數如主題、分區等。Spark Streaming 提供彈性的API，使得集成各種輸入源變得簡單高效。

2

微批次處理引擎

在輸入數據流的基礎上，開發者需要編寫處理邏輯，如過濾、轉換、聚合等操作。這些操作可以利用 Spark 提供的豐富函數庫來實現，並充分發揮 Spark 的分布式計算優勢。

3

結果輸出

最後，開發者需要定義輸出目的地，例如將處理結果寫入 HDFS、資料庫或消息隊列。Spark Streaming 提供多種輸出接收器選項，確保處理結果能夠被有效利用。

Spark Streaming的可靠性保證

1 **容錯性**

Spark Streaming **採用檢**查點機制，**能**夠在**發生故障時自動恢復，確保數據不丟失。同時，它支持多副本儲存，進一步提高了可靠性。**

2 **精確一次語意**

Spark Streaming **提供了精確一次** (Exactly-Once) **的處理語意，確保每個輸入記錄都會被精確處理一次，不會出現重複或遺漏。這對於需要高精度的應用場景非常重要。**

3 **狀態管理**

Spark Streaming **支持豐富的狀態管理機制**，**包括** Checkpoint **和** WAL (Write-Ahead Logs)，**能**夠確保狀態的可靠性和一致性，**即使出現故障也能快速恢復。**

4 **背壓控制**

Spark Streaming 內置了動態的背壓控制功**能，能**夠根據系統資源的使用情況自動調整**輸入速率，避免因處理能力不足而導致數據丟失或延遲。**

Spark Streaming實戰案例

網路流量分析

使用 Spark Streaming 處理網路流量數據，實時監控網路狀況，識別異常流量，並採取自動化措施進行調整和優化。這有助於提高網路服務的可用性和安全性。

社交媒體監控

透過 Spark Streaming 接收和分析來自社交媒體的實時數據，如推文、評論等，實時監測輿情動態，及時發現異常情況並採取應對措施。這對於提升品牌聲譽和危機管理非常有幫助。

物聯網數據處理

Spark Streaming 可以用於處理來自各種物聯網設備的海量即時數據，如車輛狀態監測、工業設備故障預警等。通過實時分析這些數據，可以提高設備利用率和運營效率。

金融交易監控

在金融領域，Spark Streaming 可用於實時監控交易活動，以快速識別潛在的欺詐行為。這不僅有助於降低風險，也能提升客戶服務的質量和響應速度。

Spark Streaming與其他框架的整合

Apache Kafka

Spark Streaming 與 Apache Kafka 緊密整合，可以直接從 Kafka 消費數據並進行處理。這種組合能夠提供可靠的端到端串流處理解決方案。

Elasticsearch

Spark Streaming 可以將處理結果直接寫入 Elasticsearch，實現實時數據的索引和分析。這種組合適用於日誌分析、應用監控等場景。

Apache Hadoop

Spark Streaming 可以與 HDFS 等 Hadoop 生態系統無縫整合，實現批次和串流數據的統一處理。這種混合架構能夠滿足企業全方位的大數據需求。

Tableau

Spark Streaming 處理後的數據可以輸出到 Tableau 等商業智能工具，實現即時數據可視化分析，為決策者提供及時有效的洞見。

結語

Apache Spark Streaming **是一個功能強大、可靠性高的串流處理框架，能**夠幫助企業快速分析和利用即時數據，以提升業務敏捷性和競爭力。通過**本文的介紹，相信您對** Spark Streaming **有了更深入的了解，並能**夠在實**際應用中靈活運用，從而推動企業的數字化轉型。如果您對** Spark Streaming **還有任何疑問，歡迎隨時與我們聯絡交流**。

工業4.0時代的大數據應用

工業4.0正在徹底改變製造業的面貌。大數據技術是其中關鍵的驅動力之一，為企業帶來巨大的機遇和挑戰。

大數據在工業4.0中的作用

1

數據採集

物聯網設備、RFID標籤、機器傳感器等技術可以在生產全過程中實時採集各種操作數據，為後續的分析和決策提供基礎。

2

數據分析

利用先進的數據挖掘、機器學習等算法，企業可以從海量的運營數據中發現隱藏的模式和規律，預測未來趨勢。

3

智能決策

基於對數據的深入分析，企業可以做出更加精準、智能的決策，在生產、供應鏈、營銷等領域實現自動化和優化。

提高生產效率

優化排程

Hadoop可以分析生產過程數據，自動調整最佳生產計畫，減少瓶頸，提高產能。

降低成本

通過預測性維護和精準規劃，Hadoop能夠有效降低維修費用和原材料浪費。

提升品質

實時監測和數據分析有助於識別品質隱患，確保產品質量。

大數據在工業4.0中的挑戰

1 數據整合

來自不同來源的數據格式、結構各異，需要進行有效的融合和清洗，以確保數據的質量和可用性。

2 分析能力

複雜的數據分析需要專業的統計、機器學習等技能，企業需要培養相關人才或尋求外部支援。

3 數據安全

海量的敏感生產數據面臨著網絡攻擊和數據洩露的風險，需要強化數據安全防護措施。

4 組織變革

大數據應用需要企業從組織架構、流程、文化等方面進行全面的數字化轉型，這需要長期的投入和努力。

展望：大數據在工業4.0中的發展趨勢

邊緣計算

將數據處理和分析能力下放到生產設備和工廠現場，提高數據處理的實時性和效率。

預測性維護

利用大數據分析預測設備故障，有助於及時維修和更換，降低設備停機時間。

個性化生產

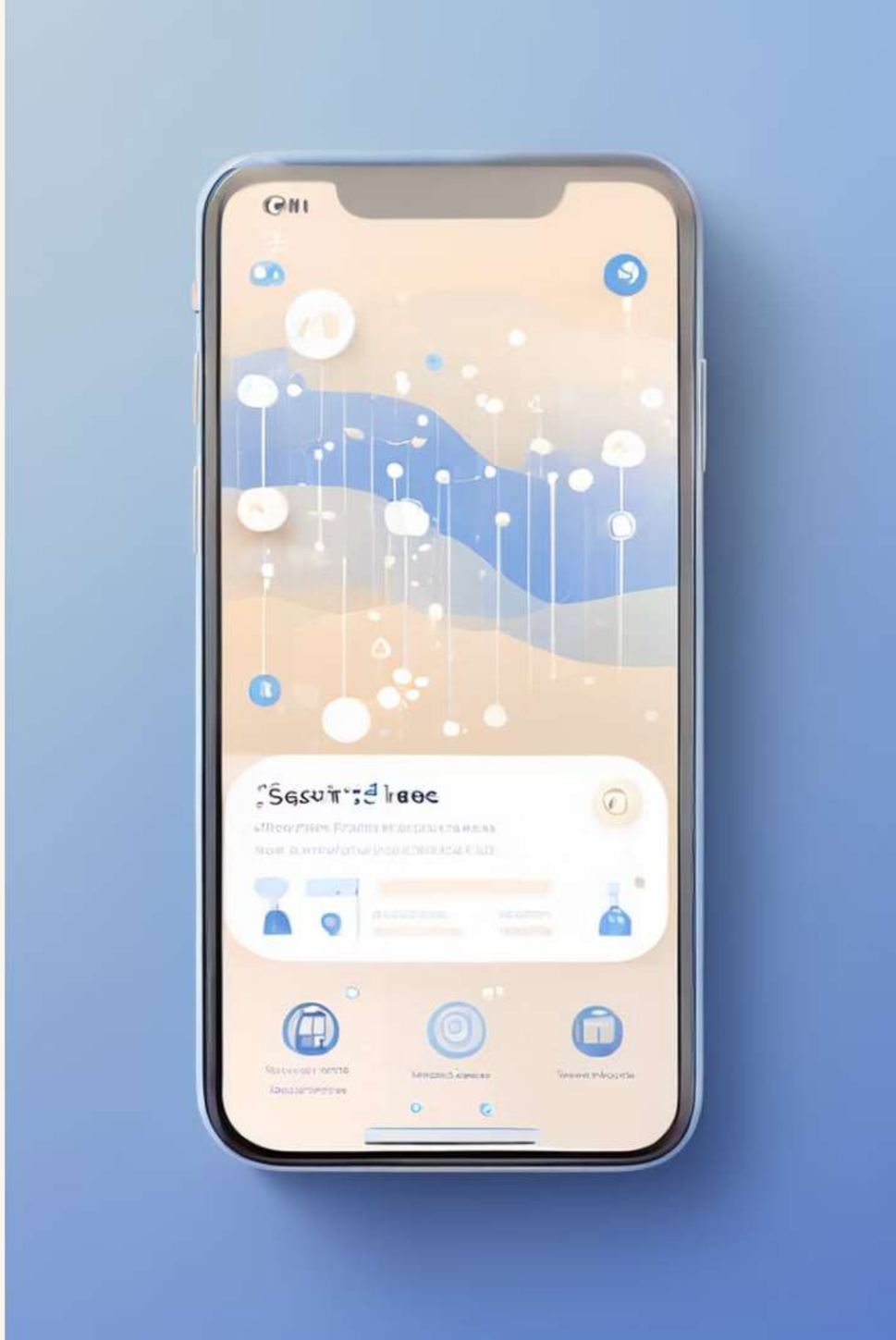
通過深入分析客戶需求，實現針對個人或小批量的定制化生產，提高客戶滿意度。

智能工廠

將大數據、物聯網、人工智能等技術深度融合，實現生產全過程的智能化和自動化。

大數據在社交軟體的應用

探討大數據如何幫助社交軟體優化使用者體驗和增加營收。



用戶行為分析

活躍度

分析用戶參與度和忠誠度的模式

興趣偏好

了解用戶對內容的喜好和反饋

社交網絡

研究用戶之間的關係和互動模式

Last 7 days

Xtra Sub

Active	New	Renewal	Expired
1010	59	8	49

Plus Sub

Active	New	Renewal	Expired
2563	164	67	148

Diamond Sub

Active	New	Renewal	Expired
308	28	10	25

Active Users

DAU	WAU	MAU	Last 10m Users
70.9 K	169 K	397 K	4.75 K

Total Users

active	paused	banned	deleted	fake
652 K	3.04 Mil	417 K	2.20 Mil	14.6 K

Purchase Diamond

334

Purchase Ticket

55

Current Online User

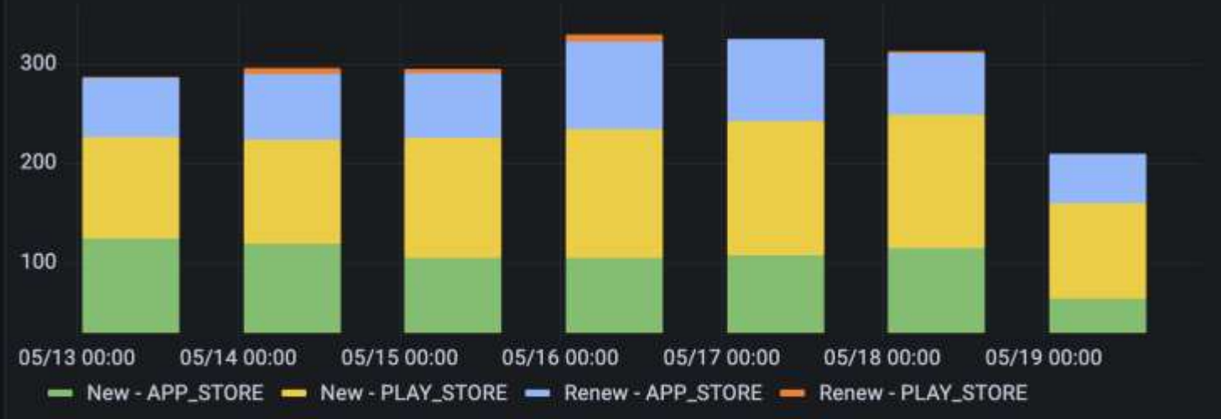
female	male
780	1144

Moderation

Image	Verification	Verification Pending	Flag	HMU Flag	Chat Flag
650	0	13851	8	16	6

New Sub / Renew Sub HIstory

New Sub / Renew Sub HIstory



US Retained User

Daily US New User

個性化內容推薦

基於興趣

根據用戶偏好推薦內容

社交影響

根據好友反應推薦內容

行為預測

預測用戶未來的興趣和行為

動態調整

持續優化演算法以提升推薦準確度

優化社交軟體營收

廣告投放

精準投放個性化廣告

內購服務

提供優質的付費增值服務

會員系統

基於用戶行為數據設計會員計劃

數據洞察

利用數據分析優化整體商業策略

clevertap_welcome_journey experiment v1

running | Make Changes | Stop Experiment

Overview | Results | Health (1)

Feb 23, 2024 — now (86 days)

Analysis Settings | Swipr-Prod | Logged-in Users | clevertap_welcome_j... | 12 metrics | last updated about 4 hours ago | Update

BASELINE: 0 use swipr server vs VARIATIONS: 1 use clevertap | DIMENSION: None | DIFFERENCE TYPE: Relative | PHASE: 1: Feb 23, 2024 — now | CUPED

Warning: Expected 2 variation ids (0, 1), but database returned 3 (0, 1, 2). Update Ids

692,445 total users >

Sample Ratio Mismatch (SRM) detected. P-value below 0.001.
Results are likely untrustworthy. See the health tab for more details.

Goal Metrics	Baseline	Variation	Chance to Win	% Change
Revenue per user	US$0.34 *US$211,725 / 626.2K*	US$0.36 *US$24,111 / 66.2K*	88.6%	↑ 7.7%
Plus Pay Rate	0.708% *4,432 / 626.2K*	0.745% *493 / 66.2K*	85.2%	↑ 5.21%
Xtra Pay Rate	0.22% *1,378 / 626.2K*	0.233% *154 / 66.2K*	73.7%	↑ 5.7%
Gem Sub Pay Rate	0.116% *728 / 626.2K*	0.118% *78 / 66.2K*	54.4%	↑ 1.33%

?

案例分析:Grindr的數據應用

用戶互動分析

了解用戶的參與度和社交網絡

1

內容趨勢監測

實時分析熱門話題和用戶情緒

2

個性化推薦

根據用戶興趣提供個性化內容推薦

3

- 推薦系統
- Spam Monitor

大數據在社交軟體的未來

更智能的推薦	結合行為分析和機器學習	提升個性化體驗
深層次洞察	整合多源數據分析用戶全貌	指導業務決策
隱私保護創新	在合規前提下保護用戶隱私	維護用戶信任