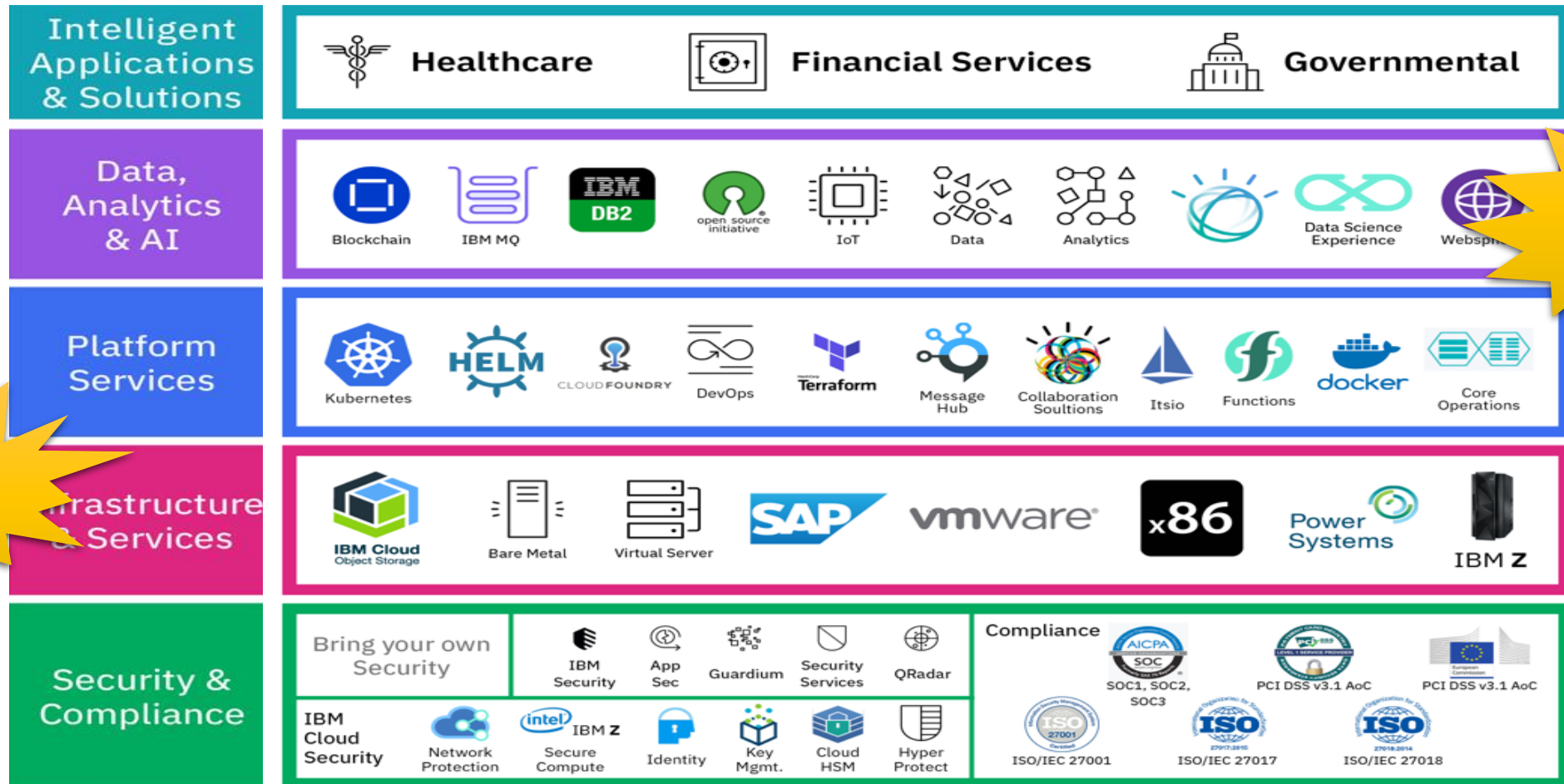




Build a data lake using object storage on IBM Cloud

Jack Wu
IBM Cloud CTP
2019/12/6

IBM Cloud Architecture – Open, Fast, Secure and AI Ready



Services Ecosystem for New AI and Cloud Native Workloads

Data Scientists & Analysts



KUBERNETES SERVICE



POWERAI



MACHINE LEARNING



KNOWLEDGE CATALOG



AI OPENSACLE



SQL QUERY



APACHE SPARK



ANALYTICS ENGINE



FUNCTIONS



EVENT STREAMS



STREAMING ANALYTICS

Backup



CLOUDANT



DB2



DB2 WAREHOUSE



POSTGRES



MYSQL



MONGODB



REDIS



ELASTICSEARCH



RABBITMQ



ETCD



RETHINKDB



JANUSGRAPH



SCYLLADB



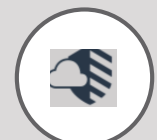
IBM Cloud
Object Storage



ACTIVITY TRACKER



KEY PROTECT



Cloud IAM

Solutions



SECURITY ADVISOR



SPECTRUM PROTECT PLUS ON IBM CLOUD



BOX



BLOCKCHAIN



CONTAINER REGISTRY



INTERNET OF THINGS PLATFORM

IBM Cloud Object Storage 的各種使用場景

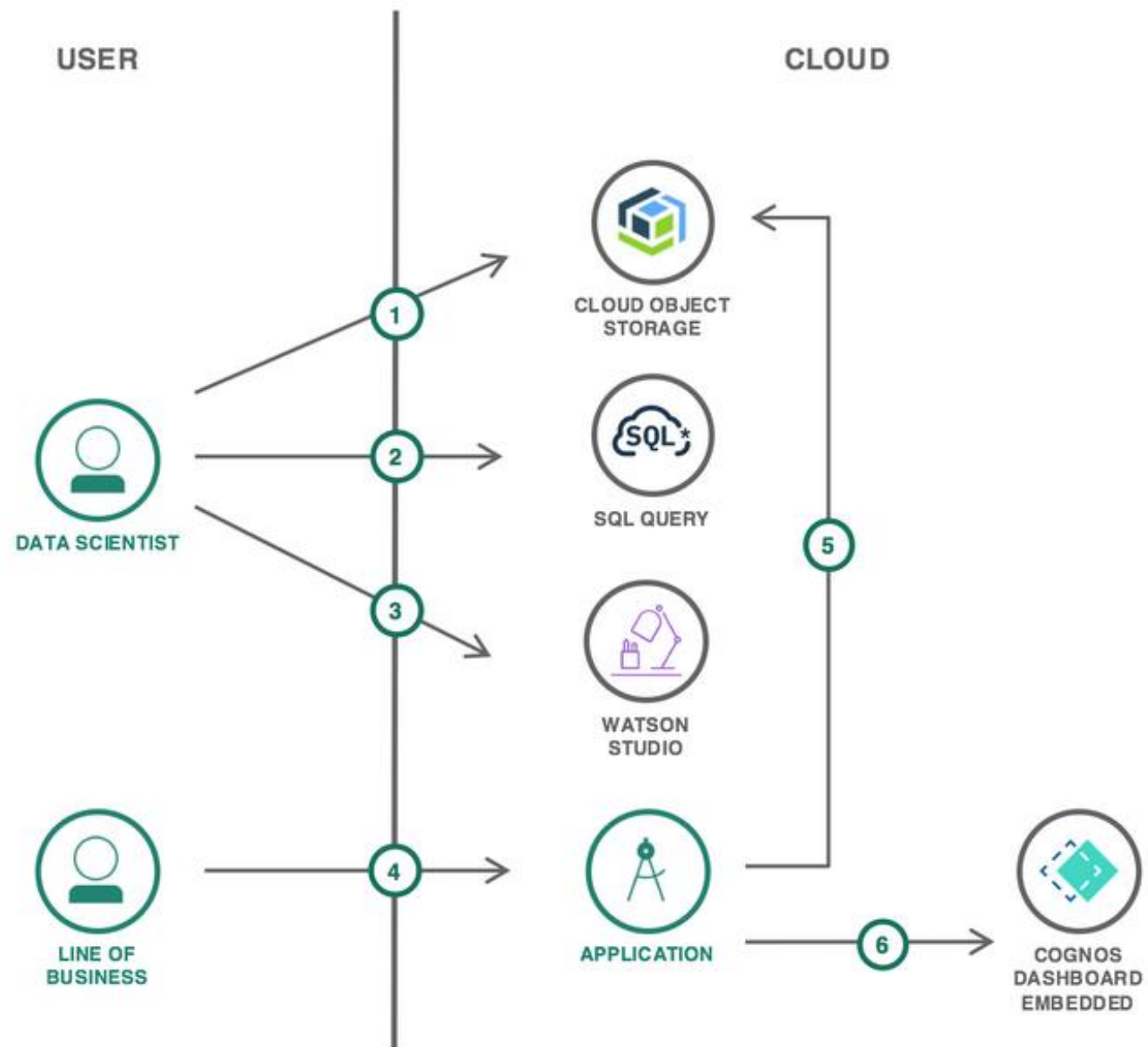
資料備份	線上歸檔	企業協作	儲存服務	資料倉儲
Backup target	Archive data	Secure sharing	Provide a service	Live content data
IBM Spectrum Protect CTERA RMAN Commvault NetBackup AltaVault Actifio Rubrik	IBM Spectrum Scale Apache Spark Apache Hadoop Analytics Avere Tiger Bridge Connector StorReduce Gorilla Video Archive Openfind Mailbase*	CTERA IBM ECM SME Hopebay ArkEase Pro	Cloud service Service Provider cloud Analytics service Video production Video surveillance Picture data service Arisan	IBM Spectrum Scale CTERA Apache Spark Apache Hadoop Analytics IBM FileNet SME Panzura Avere Nasuni



Objectives

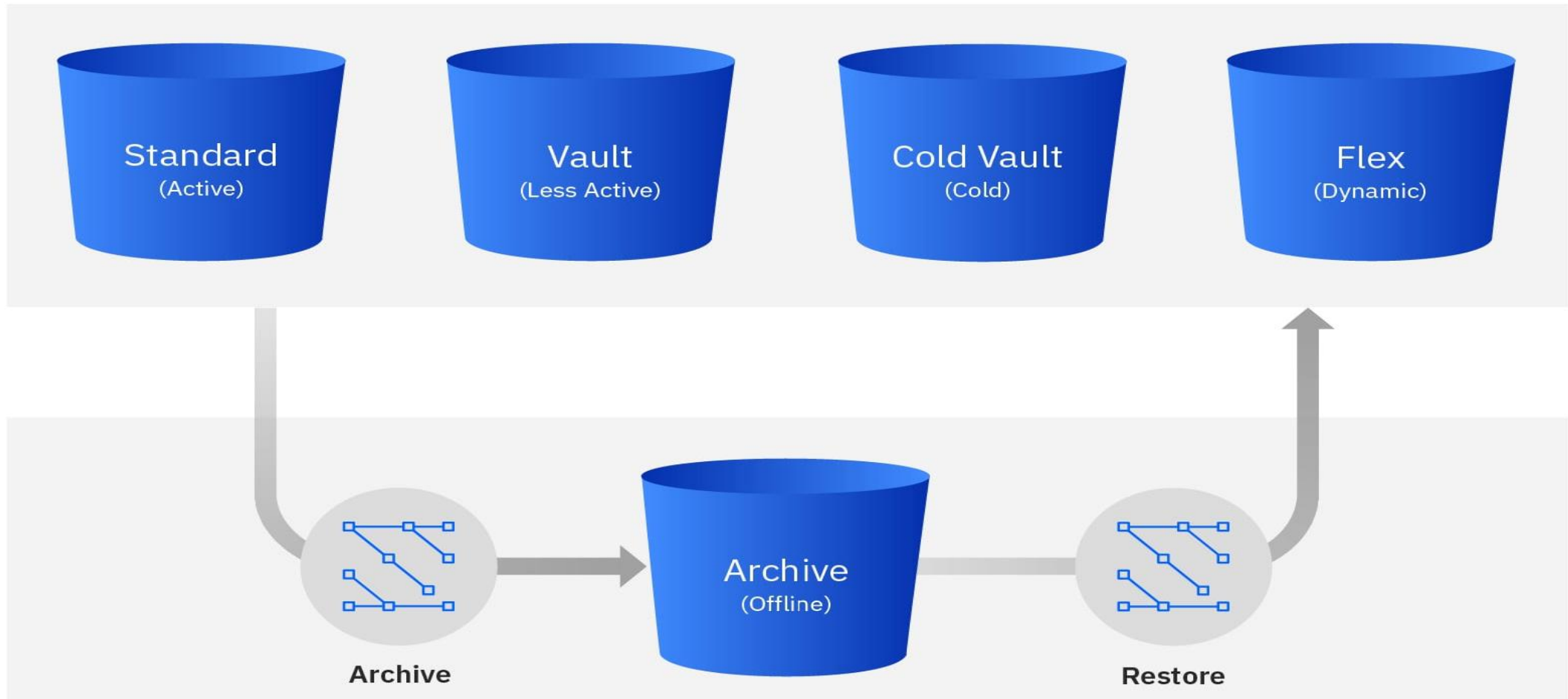
- Use Object Storage to store raw data files
- Query data directly from Object Storage using SQL Query
- Refine and analyze data in IBM Watson™ Studio
- Share data across your organization with IBM Cognos Dashboard Embedded

Architecture



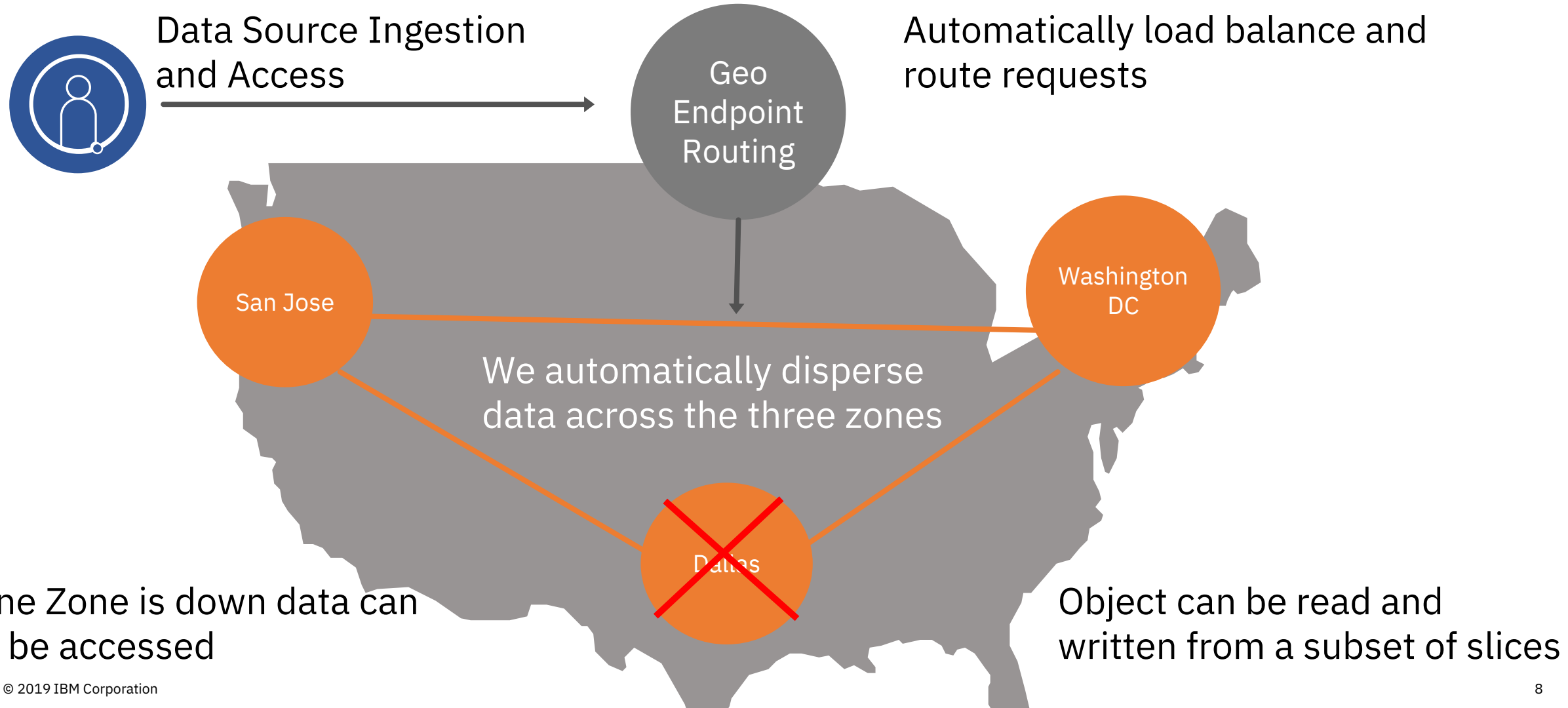
IBM Cloud Object Storage

--Low cost Archive for Long Term Data Retention



IBM Cloud Object Storage

--Cross Region Resiliency



IBM Cloud Object Storage

-- Aspera high-speed transfer

- No charge for data upload
- IBM SDKs for ease of use
- Ideal for large files and variable network connections
- Faster transfer speed than standard HTTP
- Security that starts at the point of transfer



製造業Use Case:

Big Ambitions, Big Challenges

Background

- 全球各地工廠需要PLM映像檔來進行生產，故透過FTP方式來作全球跨域傳送
- 生產需等到映像檔正確傳送完成，才可進行

Challenges

- 映像檔檔案過於龐大，無相應儲存及備援硬體
- FTP Server硬體設備成本較難估算(內部硬盤或外掛儲存設備)
- 自建的網路設備與對外流量管理極為要求，下載上傳過於緩慢，影響業務
- 工廠分散在全球，映像檔跨國存放無法集中管理

Desired Results

- 24 * 7不間斷服務
- 高可擴充性硬體
- 異地備援支援
- 最短災難復原時間
- 全球集中管理
- 穩定傳輸環境

製造業Use Case:

Big Ambitions, Big Challenges

Why IBM

- 5/20TB對外免費流量
- 跨資料中心內網傳輸免費，可做延伸機房，可配合客戶企業的全球佈局策略
- 物件儲存服務有兩大好處
 - ✓ 直接提供跨國資料保護
 - ✓ Aspera 快速上傳下載傳送服務

Final Result

- 24*7不間斷服務
- 跨國異地備援
- 高彈性FTP Server配置
- 高擴充性
- 高速上傳檔案(Aspera Upload)，搭配COS V2
- 安全防護 Security Group + Firewall + 弱掃
- 高穩定傳輸 Local Reverse Proxy FTP

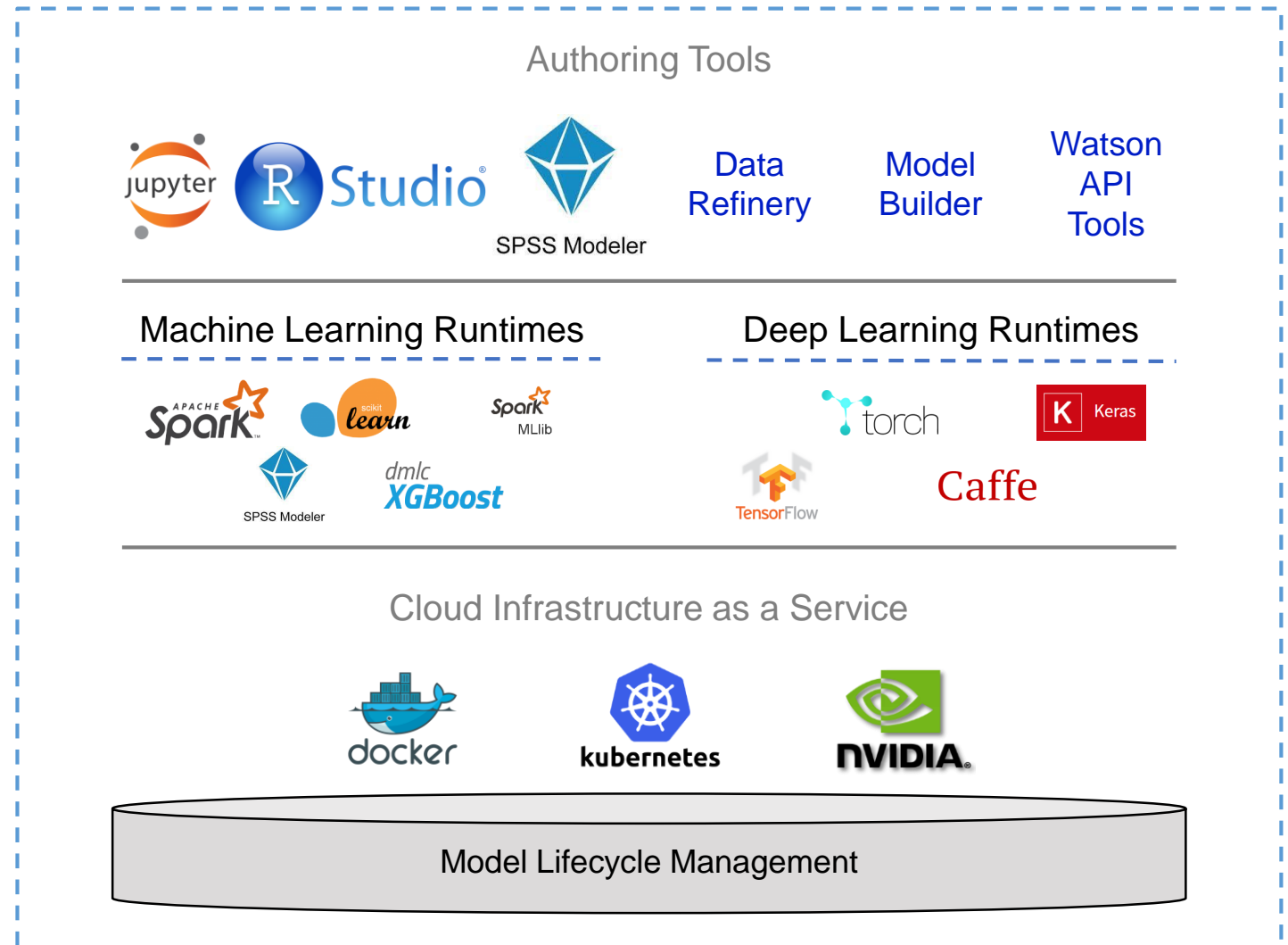
Watson Studio

為整體AI 工作流程提供最全面完整的工具組合

- 建置、協作、部署和監控
- 最佳的開放原始碼及IBM工具
- 編碼 (R, Python or Scala) 與非編碼/視覺化建模工具

- 最熱門的開源框架
- IBM 最佳框架

- 完全代管的服務
- 基於容器的資源管理
- 彈性的隨付隨用 CPU/GPU 運算資源



Watson Studio

支援目前最常見的開源框架

- 已內建整合於 Watson Studio 之中，無須安裝或配置，即可直接使用。
 - 例如：當您啟動 Jupyter Notebook 時，SparkContext 已經就緒
 - 為程式開發者提供 Jupyter Notebook；為一般使用者提供圖形使用介面或精靈
 - 可以在 Jupyter Notebook 使用 `pip install` 指令來安裝新的套件

Machine Learning Runtimes



Deep Learning Runtimes



- 在 Python 環境，已內建安裝 Pandas, NumPy, matplotlib 等程式庫



Watson Studio

The screenshot shows the IBM Watson Studio interface for a project named 'demo99'. The main area displays a Jupyter notebook with Python code for plotting a bar chart. The code includes imports for 'matplotlib.pyplot' and 'collections.Counter', and a function 'plot_tweets' that takes a list of tweets and returns a bar chart showing the distribution of tweets by country. The chart is titled 'Tweets Country Distribution based on the User Profile' and shows bars for GERMANY, MEXICO, CANADA, INDIA, and JAPAN. A blue callout box at the bottom of the notebook area contains the text 'Open Source tools – Jupyter and RStudio'.

The screenshot shows the IBM Watson Studio interface for a project named 'Insurance use case'. The main area displays the 'Car recognition' project, which includes a grid of image classes and a list of image files. The 'My Classes' section shows 49 classes, including 'AMGeneralHummers...', 'AcuraTSXSedan2012', 'AcuraIntegraTypeR2...', 'AcuraTL Sedan 2012', 'AstonMartinVBVanta...', 'AudiA8Sedan2012', and 'AstonMartinVBVanta...'. The 'All Images' section shows a list of image files, including 'BentleyMulsanneSedan2...', 'BugattiVeyron16.4Coupe...', 'BuickRegalGS2012.zip', 'BuickRainierSUV2007.zip', 'BuickVeranoSUV2012.zip', 'BuickVeranoSedan2012.zip', 'CadillacCTS-VSedan2012.zip', 'ChevroletAvalancheCrew...', and 'CadillacSRX2012.zip'. A blue callout box at the bottom of the interface contains the text 'Watson Visual Recognition – retrain Watson'.

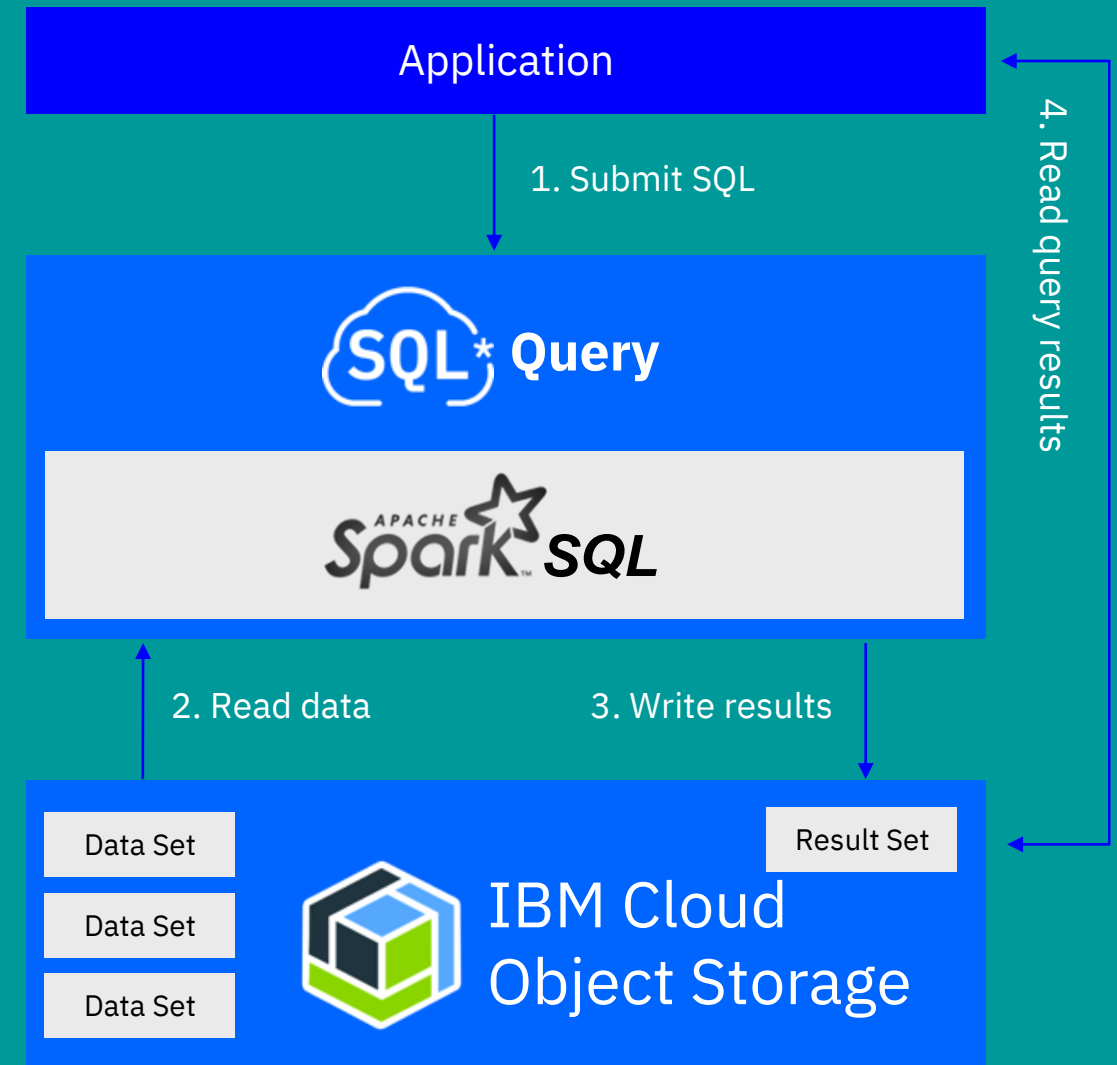
The screenshot shows the IBM Watson Studio interface for a project named 'Think99'. The main area displays the 'Environments' page, which includes a table of active environment runtimes and a section for environment definitions. The 'Active environment runtimes' table is currently empty, with the message 'you currently have no active environment runtimes'. The 'Environment definitions' section includes a table with columns for NAME, TOOL, HARDWARE CONFIGURATION, LAST MODIFIED, and ACTIONS. A blue callout box at the bottom of the interface contains the text 'Elastic and customizable compute environments'.

The screenshot shows the IBM Watson Studio interface for a project named 'Think99'. The main area displays a visual neural network flow diagram for a 'Single Convolution layer on M'. The flow starts with 'Image Data' and proceeds through 'Conv 2d', 'ReLU', and 'Pooling 2d'. From 'Pooling 2d', the flow branches into two paths: one leading to 'Flatten', 'Dense', and 'Softmax'; the other leading to 'Sigmoid Cross-E...' and 'SGD'. The 'Softmax' layer is connected to an 'Accuracy' layer. A blue callout box at the bottom of the interface contains the text 'Create ML flows and design Neural Networks visually'.

IBM SQL Query Service

IBM SQL Query is a serverless, interactive querying service to quickly and easily analyzes data in IBM Cloud Object Storage.

- Quickly submit and run SQL queries directly to Cloud Object Storage
- No setup required
- Query data where it resides
- Write results back to Cloud Object Storage
- Leverage Cloud Object Storage permissions and policies to securely access your data





Live Demo



Thank You!