La piattaforma Bigdata di Sparkle e l’evoluzione IA

Linux Day 2023
Sparkle – The company

Sparkle is a TIM-owned company that operates in a global context offering connectivity solutions and ICT services to meet the needs of multinational companies, internet service providers, OTTs, Media and Content Player, fixed and mobile phone application providers and carriers.

Sparkle owns and operates a proprietary network of over 600,000 km of land and sea cables with a widespread presence in the Mediterranean basin, Europe and America. With a direct presence in 32 countries and commercial coverage on a global scale, Sparkle can count on a workforce practically distributed all over the world.

Sparkle delivers the most sophisticated communications with the highest quality and reliability through an unmatched state-of-the-art global backbone of terrestrial fiber and international submarine cable.
Sparkle – Main Indicators

**COMPANY DATA**

- Commercial Operations in **32 Countries**
- Headcount **719**
- **600,000 km** fiber optics
- **164 PoPs** Network coverage
- PoPs via partners **>1,000**

**AT A GLANCE**

- **100% TIM Group**
- **1st International Service Provider in Italy**
- Among the **Top-10 Global Operators**
- Ranked **#5** for Seabone (AS 6762), Sparkle’s Global IP Transit Backbone *(1)*

**KEY FIGURES**

- **1,900 Customers**
- Voice Traffic **88 minutes**
- Data Traffic **77 Tbps**

---

*(1) Source Caida*
AI

AI is general purpose technology

Techniques that allow to emulate human behavior

Machine Learning
Machines that have the ability to learn without being explicitly programmed

Deep Learning
Extraction of pattern from data using Neural Networks
Value from AI Technologies: -> 3 years

Today the most relevant part of the value of AI projects is concentrated on Ads and Web Search solutions.
## Supervised vs Generative AI

<table>
<thead>
<tr>
<th>Input (A)</th>
<th>Output (B)</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>Spam/No Spam</td>
<td>Spam filtering</td>
</tr>
<tr>
<td>Image, Radar Info</td>
<td>Position of other cars</td>
<td>Self driving car</td>
</tr>
<tr>
<td>Restaurant reviews</td>
<td>Positive/Negative</td>
<td>Reputation</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Generative AI is built by using supervised learning to repeatedly predict next word**

**Adoro mangiare** (input/prompt)

- risotto con gli asparagi
- l'arrosto preparato con gli amici
- la pasta con le sarde preparata da mia madre
- fuori con gli amici

**output**

<table>
<thead>
<tr>
<th>Input (A)</th>
<th>Output (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adoro mangiare</td>
<td>risotto</td>
</tr>
<tr>
<td>Adoro mangiare risotto</td>
<td>con</td>
</tr>
<tr>
<td>Adoro mangiare risotto con asparagi</td>
<td>asparagi</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Generative AI

Scrivi un programma che effettua il training di una rete neurale con TensorFlow.
Large Language Models (LLM)

LLM can be pre-trained and fine-tuned for specific purposes;

LLM are trained to solve common language problems: text classification, question answering, document summarization, text generation, ...;

... then tailored to solve specific problems in different fields: Finance, entertainments, ...using a relatively small fields and datasets;

Benefit of LLM:
1. a single model can be used for different tasks;
2. fine-tune process requires minimal field data;
3. the performance is continuously growing with more data and parameters.

LLM projects:
✓ LaMDA;
✓ PALM (Bard);
✓ GPT-3.5/4 (175 B trainable parameters, 96 Layers; 500 B of tokens);
✓ ....

My favorite color is red

Input Layer

Hidden Layers

Output Layer

3 main kinds of LLM:
- Generic (next token);
- Instruction tuned;
- Dialog tuned;

Reinforce learning:
Es Proximal Policy Optimization (PPO)
AI Risks/Opportunities

✓ AI (do anything a human can do) is still decades away;

✓ AI develops gradually and the «hard take off» scenario is not realistic;

✓ AI is an important piece of solution to the real existential risks to humanity (next pandemic, climate change, ...);

✓ AI problems: bias, fairness, accuracy.....but the technology improve quickly;

✓ AI will disrupt many occupations.
The rediscovery of Neural Networks

1. Big Data
   - Larger datasets
   - Easier collection and storage
   - Centralized governance

2. Hardware
   - GPUs
   - Parallel architectures

3. Software
   - Improved techniques
   - New modes
   - Toolboxes
From DWH to Big Data

The value of the evolution of Big Data has been the unification of the data of the entire corporate group, which is geographically distributed throughout the world and has various Legal Entities.

In the past decades Sparkle developed different DWHs for separate data analysis:
- **STS** – for elementary Voice Traffic
- **WWW** – for Voice and SMS aggregated and rated traffic
- **QUID** – for Voice and IP Quality

More recently the project One Customer View (OCV) realized a new unitary vision for customer data, based on the gathering from multiple sources and introducing the presentation on mobile devices (MobApp)

Big Data has been introduced to go beyond the single DWH and move towards the Enterprise Data Integration. Different streams have been defined in order to address the needs of different processes:
- Economics Performance & Predictions
- Products Dashboards
- Traffic and Quality Analisys
- Process Monitoring
- Proactive Monitoring
Big Data – Data Platform on Premises

- **ETL**
  - Internal and external data sources
  - ETL
- **CDP Private Cloud Base**
- **Open LDAP / Kerberos**
- **Nutanix / Acropolis**
- **Tableau**
- **Oracle**
- **Trino Query Engine**
- **Cloud Data Services (K8s cluster)**
- **RA/RCEM – GS Advanced Analytics**
- **API**
- **Apps**
Big Data – Data Platform on Premises / Open Sources

Data Storage
- Apache HDFS
- Ozone

Data Ingestion & Streaming
- Apache Kafka
- Apache Spark
- Spring Boot

Data Processing
- Apache Spark
- jupyterLab
- Zeppelin

Data Governance
- Atlas

DWH
- Apache Hive

Workload Management
- Apache YARN
- Kubernetes

API
- Spring
Big Data – Open source data platform

Data Storage
- S3
- Ozone/HDFS
- NOSQL

Data Sources
- Kafka
- Data Ingestion/Collections

Data Processing
- Spark SQL
- Hive
- presto
- jupyterLab
- Apache Airflow
- Apache Beam

Security Layer

Data Visualization

- Intelligent Agents
- BI
- Recommendation
- Real-time Dashboard

Data Ingestion/Collections

Data Analytics
- DWH
- NLP Predictive Statistical

Processing
- Batch/Realtime

Intelligent Agents

Reccomandations

Real-time Dashboard

JSON

Batch

Realtime

Dashboard

Data Visualization

La piattaforma Bigdata di Sparkle e l’evoluzione IA
Palermo, Oct 2023
Big Data – Use cases implemented

- **Revenues**
- **Volumes**
- **Costs**
- **Costs Reconciliation**
- **Revenues Reconciliation**
- **POP Profitability**
- **Business View for Services**
- **Margin**
- **Billing Active/Passive Cycle**
- **QoS IP&Data**

Profitability Performance
Big Data – Use cases implemented (ML) : Clustering

Using K-means implemented in sklearn library
Big Data – Use cases implemented (ML) : Customer Churn

Pycaret is an open-source low code ML library in Python that automates ML workflow.

Comparing model performances
Big Data – Improve productivity using GenAI {1/2}

```scala
val conf = new SparkConf().setAppName(appName);
val sc = new SparkContext(conf);
val spark = SparkSession.builder.config(conf).getOrCreate()

val schema = new StructType()
.add("timestamp_cdr", "long").
.add("type", "integer").
.add("observation-domain", "integer").
.add("observation-point-id", "long").
....

val df = spark.readStream
  .option("sep", ",")
  .schema(s1chema)
  .csv(inputPath)
  .withColumn("file_name", input_file_name())
val dfj = df.join(dfDecodeOp, df("tcap-local-operation-code") === dfDecodeOp("code")."leftouter")

val queryDF = dfj.writeStream
  .foreachBatch ((batchDF: DataFrame, batchId: Long) =>
    batchDF.persist()
    batchDF.groupBy(
      date_format(col("timestamp_cdr")).cast("bigint") / 1000).cast("timestamp")."yyyy.MM-dd HH", alias("Time")
      local_mode="true",
    col("mpl3-opc")
      ).
          .when(col("tcap-local-operation-code") isNull, value = "No data").when(col("operation") isNull, value = "No data").otherwise(col("operation")).as("tcap-local-operation-code")
          .as("tcap-message-type")
          )
          .count().withColumnRenamed("count", "ALL").coalesce(1)
          .write.mode("append")
          .option("compression", "gzip")
          .option("header", true)
          .csv(outputPath)
    batchDF.unpersist()
  )
  .trigger(Trigger.ProcessingTime(trigger ProcessingTime))
  .option("checkpointLocation", checkPointPath)
  .start()

val queryDFSStat = dfj.writeStream
  .foreachBatch ((batchDF: DataFrame, batchId: Long) =>
    batchDF.persist()
      batchDF.groupBy(
        date_format.current_timestamp(), "yyyy-MM-dd")
        .as("date_msu")
        .as("timestamp")
        .as("timestamp")
        .as("timestamp")
        .functions.reverse(functions.split(col("file_name"), "/")().as("file_name"))
        .when(col("tcap-local-operation-code") isNull, value = "No data").when(col("operation") isNull, value = "No data").otherwise(col("operation")).as("tcap-local-operation")
      )
      )
      .count().coalesce(1)
      .write.mode("append")
      .option("header", true)
      .csv(statPath)
    batchDF.unpersist()
  )
  .trigger(Trigger.ProcessingTime(trigger ProcessingTime))
  .option("checkpointLocation", statCheckPointPath)
  .start()
```
Big Data – Improve productivity using GenAI {2/2}

scrivi un job spark in scala che definisce uno stream su un folder contenente file csv

```scala
val spark = SparkSession
  .builder()
  .appName("CSVFileStreamExample")
  .getOrCreate()

// Defini lo schema dei dati CSV
val schema = StructType(List(
  StructField("col1", StringType, true),
  StructField("col2", IntegerType, true)
))

// Crea uno stream su questa directory
val stream = spark.readStream
  .schema(schema)
  .option("header", "true")  // Se la prima riga contiene l'intestazione
  .csv(csvDirectory)

....

// Esegui la count su "col1" e scrivi i risultati in una directory
val countCol1Stream = stream
  .groupBy("col1")
  .count()
  .writeStream
  .outputMode("complete")  // Usa "complete" in modo da aggiornare la count completa a ogni micro-batch
  .format("console")  // Puoi modificare il formato di output a seconda delle esigenze
  .option("checkpointLocation", "path/to/checkpoint1")
  .start("path/to/output1")

// Esegui la count su "col2" e scrivi i risultati in un'altra directory
val countCol2Stream = stream
  .groupBy("col2")
  .count()
  .writeStream
  .outputMode("complete")  // Usa "complete" in modo da aggiornare la count completa a ogni micro-batch
  .format("console")  // Puoi modificare il formato di output a seconda delle esigenze
  .option("checkpointLocation", "path/to/checkpoint2")
  .start("path/to/output2")

// Attendi fino a quando entrambi i processi di streaming non vengono terminati
countCol1Stream.awaitTermination()
countCol2Stream.awaitTermination()
```

dichiara il separatore all'interno dei csv è il "

```scala
....
// Crea uno stream su questa directory, specificando il separatore "\""
val stream = spark.readStream
  .schema(schema)
  .option("header", "true")  // Se la prima riga contiene l'intestazione
  .option("delimiter", ")")  // Specifica il separatore
  .csv(csvDirectory)
```
Big Data – Data Platform @ GCP

- Reduces operation costs using managed services;
- High productivity and quick collaboration;
- Reliability and high scalability.
allows teams to define and manage infrastructure using version control, which makes it easier for multiple people to collaborate and work on the same codebase;

allows to define infrastructure using a high-level configuration language.
Big Data – Network and security on GCP

- 6 security context (VPC):
- 2 x frontend (prod, noprod);
- 2 x backend (prod, noprod);
- 1 management;
- 1 external (on-promise connectivity)

Fortigate group to manage FW rules between external context and internal contexts
Next Step – Big Data Transformation

Advantages of the Hybrid cloud: scalability, security and flexibility

An important step in becoming a data-driven organization will be the adoption of a Unified Data Fabric, which will transform the Private Cloud architecture of Big Data into a Hybrid Cloud architecture.

The following functions will be centralized and simplified:

- Data Catalog
- Data governance
- Data security
- Centralized user provisioning
- Centralized environments provisioning
Next Step – Lakehouse Trend

"Data Lakehouses integrate and unify on a single platform the capabilities of data warehouses and data lakes, aiming to support AI, BI, ML and data engineering”

Gartner Research January 2022

1. Hybrid Data Platform
   - A single data repository requires less time and budget to administer

2. Data Lakehouse
   - Direct access to broader dataset for BI tools and ML analysis

3. Advanced Analytics
   - Advanced Analytics like Real time & Streaming Analytics
   - Simplified data governance due to having a single control point
   - Less data redundancy
   - ACID-compliant transaction support.
Next Step – Advanced Analytics

**Mission:** give voice to data!!

**Correlation Analysis and Pattern identification**
- Clustering and Correlation ML algorithms to identify *event correlation and patterns*
- Detect *hidden phenomena*, such as abandonment proneness

**Insights & predictions**
- *Ensemble of advanced ML models* for defining business insights
- Data Classification e Data Regression
- GeoAnalysis

**Business Value**
Support on strategic business choices to different company and market players

**Speed & Quality**
Ready to go! Increased time to market with high quality and efficiency

**AI&ML Capabilities**
Hidden phenomena and unknown cases identification
## AI Networks: Use Cases

<table>
<thead>
<tr>
<th>Process autom. tools</th>
<th>Service Problem Management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Repetitive tasks automation</strong></td>
<td><strong>Incident Automation – smart view e-mail Mgmnt</strong></td>
</tr>
<tr>
<td>Automation of repetitive tasks on the BB IP network (software updates, tests, massive configurations, security patches)</td>
<td>Improve the management of customer/supplier communications through automation capabilities.</td>
</tr>
<tr>
<td><strong>SLA Mng.</strong></td>
<td><strong>Proactive Customer Notification</strong></td>
</tr>
<tr>
<td><strong>SLA Management</strong></td>
<td>Due to a Major Fault, it is required to proactively inform the customer of the failure of the internal network</td>
</tr>
<tr>
<td><strong>Fault &amp; Perf. Mng</strong></td>
<td><strong>RFO Management</strong></td>
</tr>
<tr>
<td><strong>IP Network-Monitoring</strong></td>
<td>Production of a Report that contains ticket data (e.g. automatic management of RFO Customer requests)</td>
</tr>
<tr>
<td><strong>Closed-loop automation</strong></td>
<td><strong>Network AIOps</strong></td>
</tr>
<tr>
<td><strong>Automatic Remediation</strong></td>
<td>Fault prediction, capture and processing of weak signals on the network in order to prevent faults</td>
</tr>
<tr>
<td><strong>Fault prevention</strong></td>
<td><strong>SPANLOSS Mgmnt:</strong> Automatic adjustment of the transmission laser power due to high signal attenuation</td>
</tr>
<tr>
<td></td>
<td><strong>Automatic Rerouting</strong> of transmission circuits in case of Major Fault</td>
</tr>
</tbody>
</table>
La piattaforma Bigdata di Sparkle e l’evoluzione IA

AI Networks: Architecture

IBM Watson AIOps Management Platform

- Event Consolidation
- Event Deduplication
- Event Analytics
- AutomationHub (RunBooks/PlayBooks)

IBM Intelligent Network Performance Platform (NPM)

- Data Collection Scalable & Resilient
- Data Insight Graph QX
- Recall/reports in context REST API

IBM CP for AIOps (on-prem)

IBM Cloud Platform (SaaS)

ML-based events/alerts

In-context drill down per device/interface metric and flow-based analytics

IBM Cloud Platform (SaaS)

IBM Watson platform

Extract actionable data from unstructured data

IBM Nework Performance SevOne (on-prem)

Solution is composed of specialized elements that interact each other in a flexible way in order to follow an agile approach in the implementation/direction of the various use case.

La piattaforma Bigdata di Sparkle e l’evoluzione IA

Palermo, Oct 2023

27
GRAZIE