ADVANCED ANALYTICS

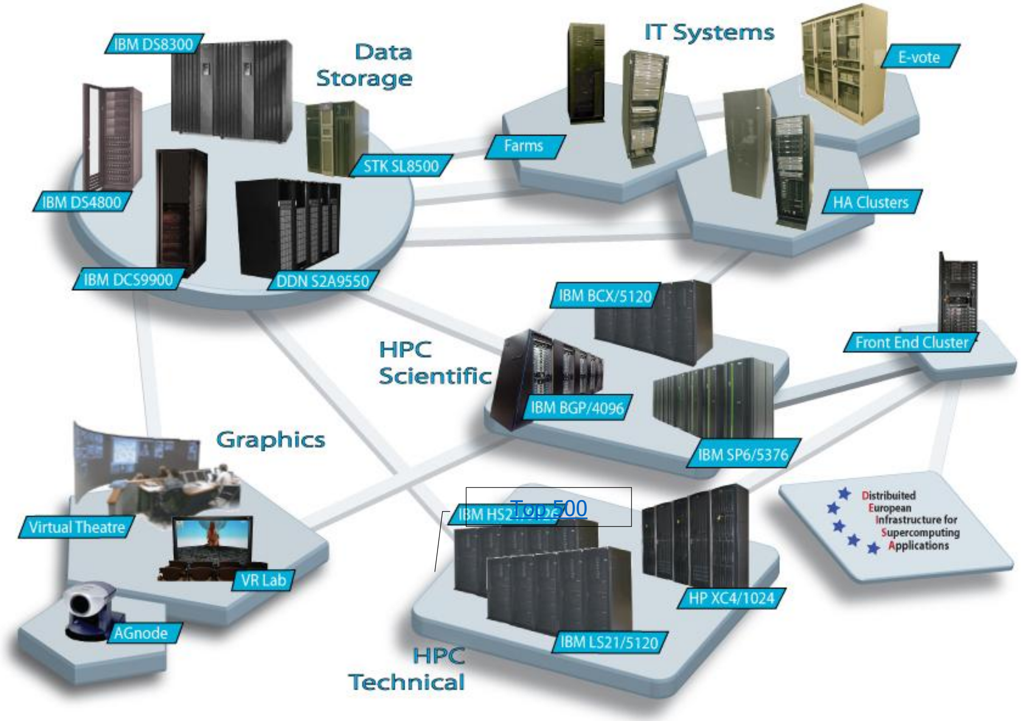
Stefano Roselli s.roselli@cineca.it

CINECA

Il Cineca è un Consorzio Interuniversitario senza scopo di lucro al servizio del sistema accademico nazionale istituito nel 1969.
La missione è “promuovere l’utilizzo dei più avanzati sistemi di **elaborazione dell’informazione** a favore della ricerca scientifica e tecnologica, pubblica e privata”, e al trasferimento tecnologico alle imprese e alla Pubblica Amministrazione.

Fanno parte del Consorzio:
- MIUR
- 70 università
- 4 Enti di Ricerca

Circa 900 dipendenti con sedi
a Bologna, Milano e Roma



PRACE

BDV

Il Laboratorio Big Data & Analytics

Il Laboratorio di Big Data & Analytics è una iniziativa di CINECA, nel campo della High Performance Analytics per promuovere la sua diffusione e aiutare i decisori aziendali e i professionisti ICT a comprendere le strategie, le potenzialità e le tecnologie dei Big Data e delle tecniche di Data Mining.

PIATTAFORME SOFTWARE:
- IBM Big Insights
- Hortonworks Data Platform

ARCHITETTURE:
- Data Streaming Analysis
- **Large Scale Machine Learning**

TECNOLOGIE:
- Hadoop (HDFS, MapReduce), YARN
- Spark SQL, Hive e HBase
- Storm, Spark Streaming
- Kafka & MQTT
- Spark R e Distributed R
- Librerie: Spark MLLIB, H2O

INFRASTRUTTURA:

HPC IBM NeXtScale server appositamente progettata per i casi di calcolo “data-intensive”:

- 70 nodi IBM NeXtScale con interconnessione a 56 GB/sec
- Intel Ivy Bridge 20 core per nodo, 1480 core in totale
- 128 GB RAM per nodo
- 40 TB SSD locale al nodo, 16 PB di storage in linea

Advanced Analytics

Gli **Advanced Analytics** sono applicazioni informatiche che usano metodi matematici e statistici su sistemi computazionali altamente scalabili per **estrarre valore dai dati**, come trovare schemi ricorrenti (patterns), raggruppamenti (clusters) e relazioni nei dati (rules) per predire futuri comportamenti o scenari, fornendo anche raccomandazioni.

Analisi Predittiva nel CBM

La manutenzione predittiva nella Condition-based maintenance (CBM), si focalizza sull'individuare la probabilità di guasti **prima** che avvengano.

L'applicazione di **machine learning** per predire situazioni di probabili guasti si basa sul *costruire un modello usando dati storici e* ***addestrarlo*** *con casi noti, per essere in grado di identificare o classificare situazioni di potenziali guasti e non.* Il modello dovrà essere **validato** usando dati reali di test prima di applicarlo. La validazione fornisce una indicazione (matrice di confusione) sull'attendibilità del modello individuando i veri positivi, i veri negativi, falsi positivi e i falsi negativi.

Machine Learning

Supervised learning
Il sistema apprende da un insieme di esperienza già classificate

Unsupervised learning
Non si hanno casi da cui il sistema può apprendere

Algoritmi Predittivi

Categorical Target Variable:

- **Decision Tree**
- **Random Forest**
- **Neural Networks**
- Support Vector Machines
- K-Neraest Neighbor
- **Logistic Regression**
- **Gradient Boosting Machine**

Continuos Target Variable:

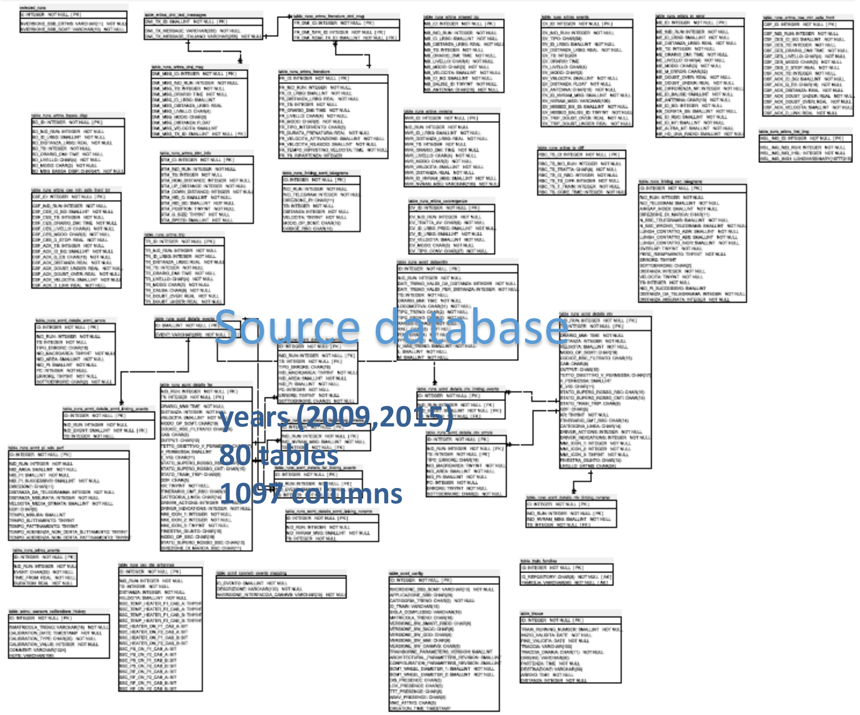
- Linear Regression
- Generalized Linear Model

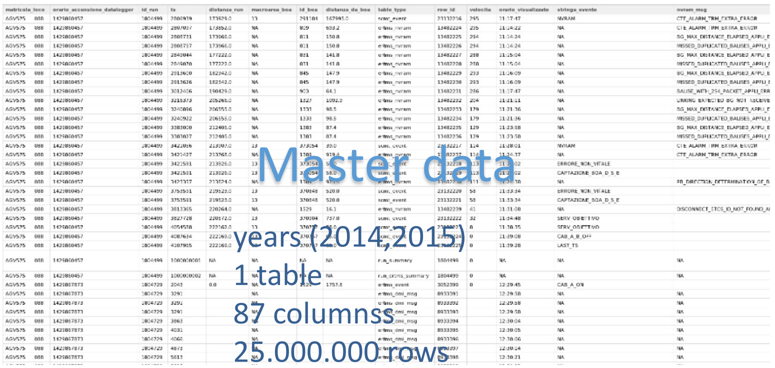
Algoritmi Descrittivi

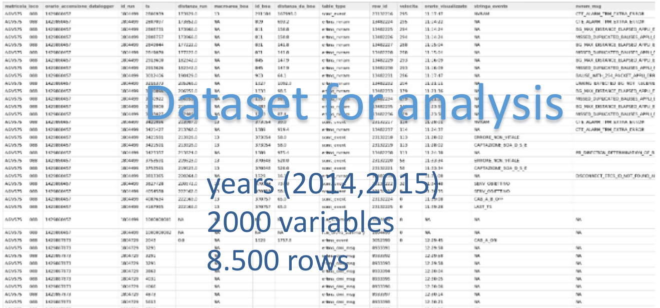
- **Clustering (K-Means)**
- Hidden Markov Chains
- Principal Component Analysis (PCA)
- Self-Organizing Maps (SOM)
- Modelli Causali

Caso Alstom: Data Preparation

- ✓ **Individuate le tabelle e le colonne del database utili all'analisi**
- ✓ **Creato il Master Data con le variabili necessarie all'analisi**
- ✓ **Creato il Dataset per gli algoritmi di Machine Learning**







Caso Alstom: Analisi Predittiva

Obiettivi

1) Valutare in modo automatico se una segnalazione di guasto attivata dal sistema di monitoraggio, sia effettivamente da segnalare all'area di manutenzione

Tecniche di Machine Learning utilizzate

- Decision Tree
- Random Forest
- Neural Networks
- Gradient Boosting Machine

Variabili per ogni evento osservato: tot. 300 var. delle 2.000 disponibili

- logs eventi diagnistica sw AV (Alta Velocità)
- logs eventi diagnistica sw SCMT (Sistema Controllo Marcia Treno)
- eventi rilevanti delle corse AV
- eventi rilevanti delle corse SCMT
- eventi relativi alla odometria

last_run_id	ISSUE	1__CAB_A_ON	1__SERV_TETTO	1__EB_ANTENNA_SWITCH	......	2__CAB_A_ON	2__SERV_TETTO	2__EB_ANTENNA_SWITCH	....	3__CAB_A_ON	3__SERV_TETTO	3__EB_ANTENNA_SWITCH	vero/falso
1559453	1	30	50	25		6	41	28		29	51	17	0
1561388	2	17	55	16		23	33	28		12	52	25	1
1561966	1	13	67	11		30	50	26		5	41	27	1
1593270	3	15	67	14		29	45	24		7	45	28	1
1656659	2	16	72	30		27	43	21		9	49	32	0
1656661	2	16	72	32		27	43	21		9	49	32	0
1656676	1	21	72	47		27	43	21		9	49	32	0
1699514	1	19	97	12		22	80	13		20	83	15	1
1704569	1	13	66	15		16	103	14		24	56	15	1
1748299	1	23	78	10		26	80	15		14	40	14	0
1783005	1	32	42	16		17	61	24		15	108	14	0
1817617	1	27	67	13		21	42	11		16	52	10	0
1653170	1	20	35	18		37	66	28		39	32	17	1
1658885	1	23	61	12		18	30	14		43	69	36	1

all'evento

500 Km dall'evento

1000 Km dall'evento

Caso Alstom: risultato ottenuto

Dal risultato ottenuto, emerge la possibilità di ridurre del 25% le false segnalazioni di guasti, che potrebbe tradursi in una riduzione dell’impiego del personale di manutenzione.

Grazie per l’attenzione

Stefano Roselli   s.roselli@cineca.it