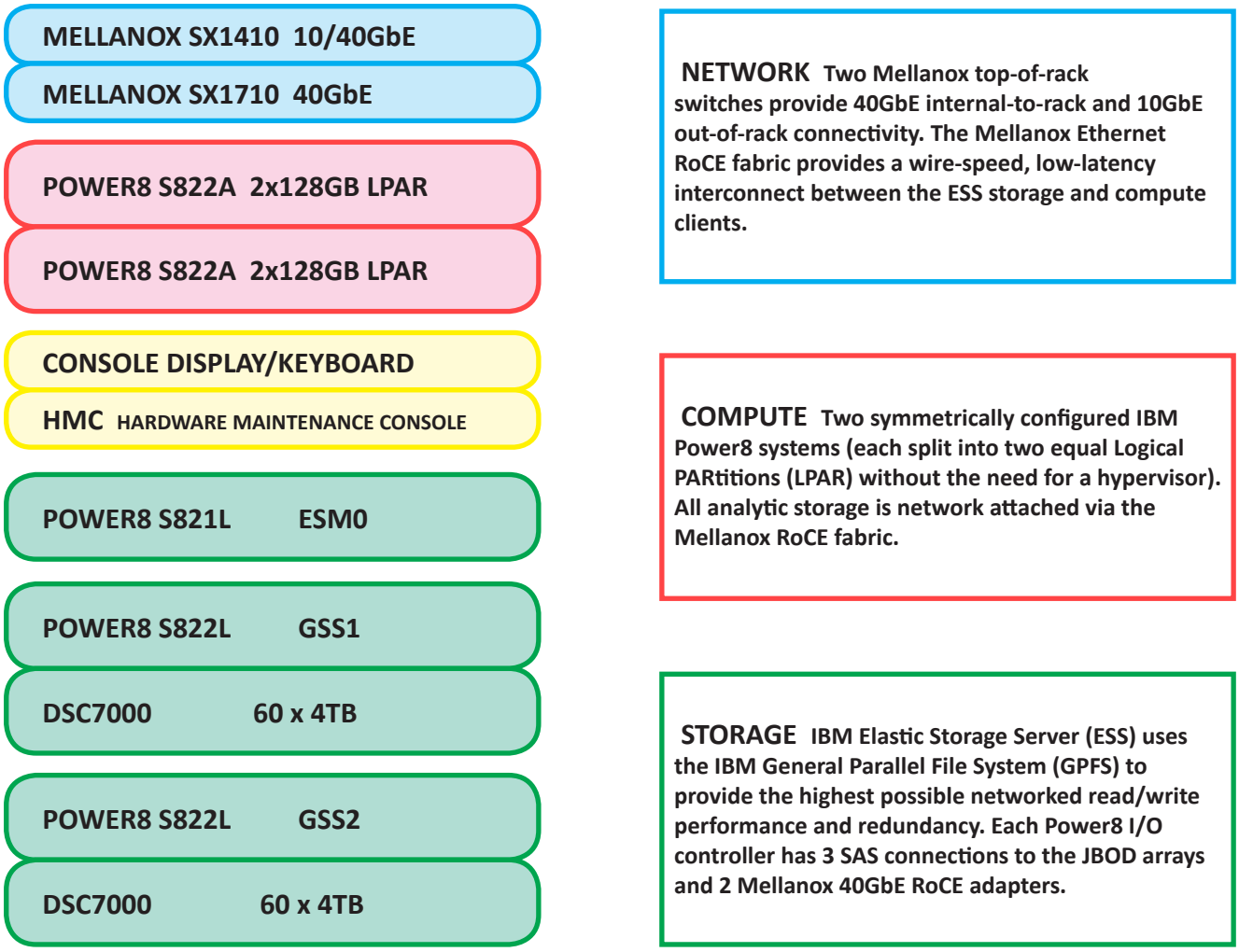
The Institute of Next Generation Computer Systems (ITng) and the Computer Science Department at North Carolina State University have developed a business process coupled with proprietary and open-source software to generate a cognitive computing platform capable of fact extraction and result scoring.

Introduction

This “Big Data” computing platform provides elements necessary to achieve high speed processing of large volumes of a variety of types data and then offer enhanced decision making services.

Development of these analytical tools is daunting given the number of software platforms, infrastructure requirements, types of questions posed and diversity of data required to generate a system of sufficient accuracy to provide relevant facts necessary for decision-making. This complex task of question disambiguation, fact identification and analysis means that the system must be capable of ingesting common forms of information, perform a progressive series of filtering and offer analytics for user-specific consolidation of relevant information necessary to make an informed decision. Functionally, this cognitive platform needs to convert information to knowledge while optimizing computing infrastructure. This process of coupling critical thinking processes with quantitative, and more importantly, qualitative information has proven to be very effective in several proof of concept experiments performed within various business verticals. The first commercial application of this technology is being deployed at a pharmaceutical company and is described below.

HARDWARE



SOFTWARE

OVERVIEW The objective of this process is to turn large quantities of vague, unstructured data into explicit structured tables that can be used to gain insight into a set of questions.

GATHER Crawl the web to get unstructured data. Web pages in an initial URL are scraped and parsed into an index-able format and stored for later use. Any relevant structured data (in existing databases) is collected and stored for use in the restructuring step.

FILTER Because web crawling is unspecific, there is much noise in the initial dataset. Data are whittled using mind-maps or PESTLE (Political, Economical, Social, Technical, Legal, Environmental) trees.

ANNOTATE A second dataset culling using regular expression matching to pull out a word from a sentence or a sentence from a paragraph to find specific information relevant to the questions being asked.

RESTRUCTURE Export data to a structured format (usually CSV or JSON) to a SQL database. Typically, each annotation is given its own table along with tables of structured data gathered in the first step.

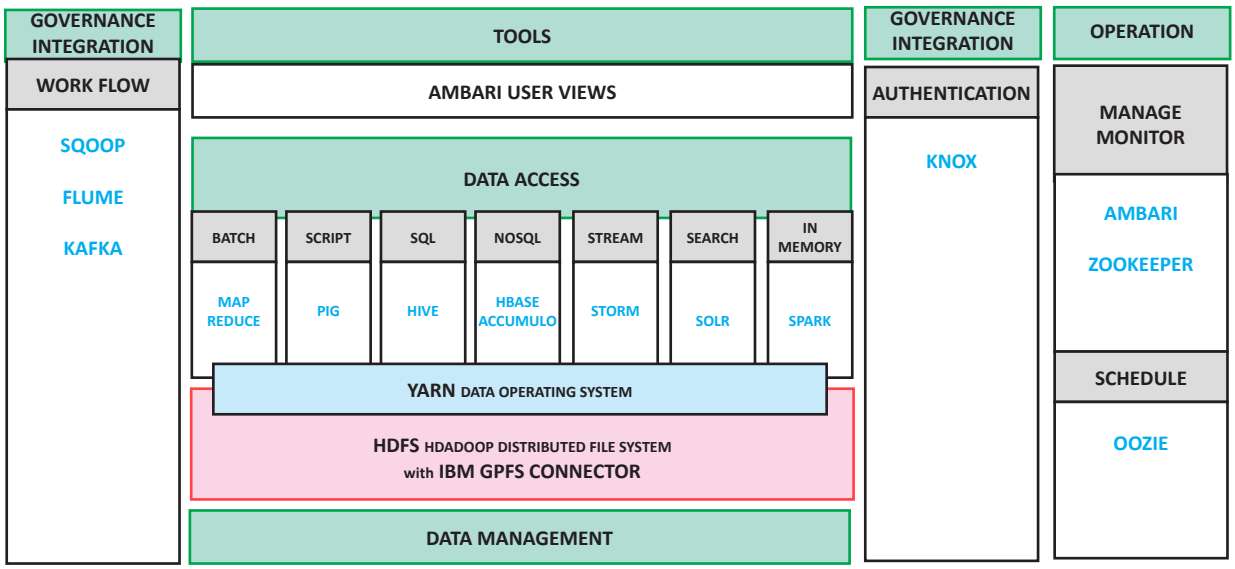
VISUALIZE Now that data are in an easily-managed format, the relationships between annotations and structured data can be graphically presented in truth tables with adjustable weighted values as answers to the questions.

PROPITIATORY SOLUTION - IBM WATSON EXPLORER

IBM Watson Explorer Advanced / IBM Watson Knowledge Studio Premium

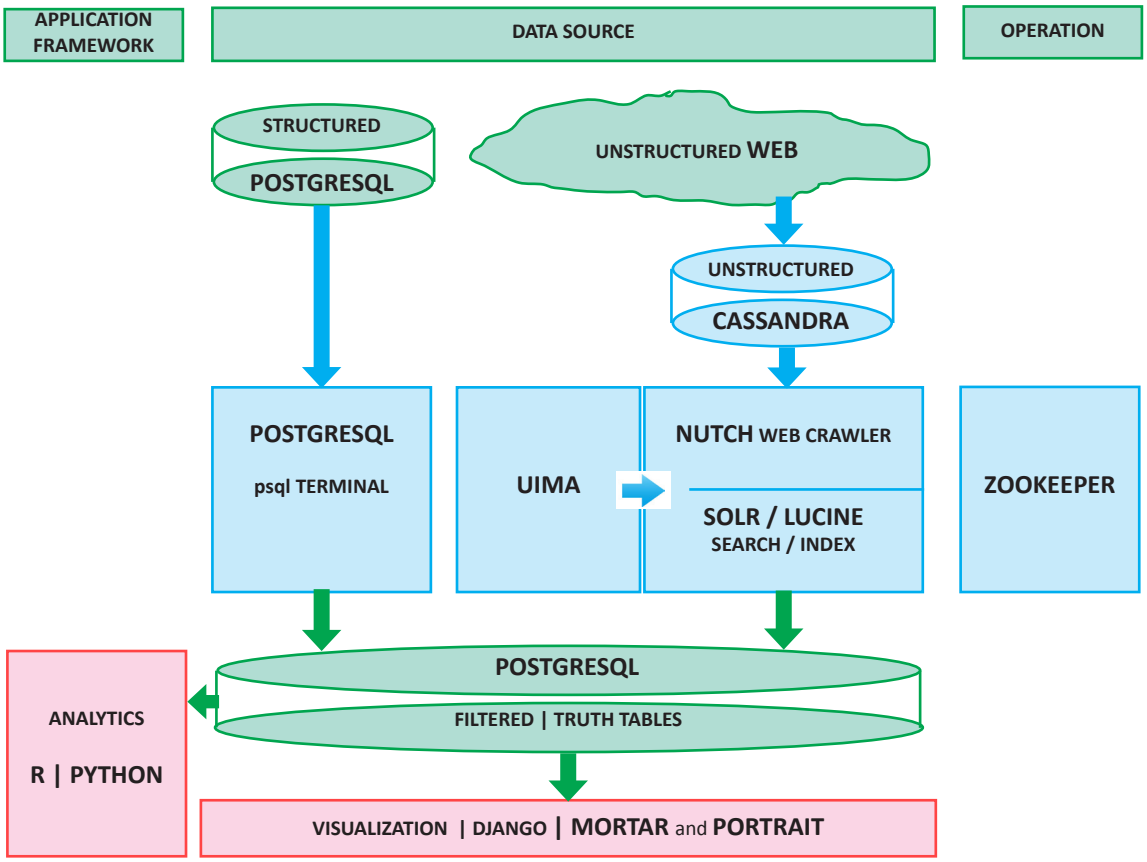
OPEN SOURCE SOLUTION - HORTONWORKS HDP

HORTONWORKS DATA PLATFORM Apache™ Hadoop® distribution based on a centralized architecture (YARN)



OPEN SOURCE SOLUTION - CUSTOM PORTFOLIO

ITng TARGET SOLUTION



STRUCTURED DATA enters via import of external databases

UNSTRUCTURED DATA obtained by NUTCH web crawler and processing via SOLR/LUCINE controlled by UIMA

UIMA (Unstructured Information Management Applications) used for analyzing data and extracting relevant entities using OpenNLP (Natural Language Processing UIMA component)

NUTCH is able to use CASSANDRA as a storage back-end and SOLR as an indexer

POSTGRESQL database holds structured (filtered data and/or truth tables) for analysis via available or custom R and/or PYTHON tools

Using CASSANDRA instead of other NUTCH back-ends (like HBASE) to avoid installing the entire HADOOP infrastructure, for ease of expansion and flexible query language (psql)

VISUALIZATION and ANALYTICS by ITng developed DJANGO web applications MORTAR and PORTRAIT

MORTAR is used create trees and mind-maps used for first-round data filtering

PORTRAIT uses DJANGO SQL EXPLORER cerate data visualizations

ENABLING TECHNOLOGY

MELLANOX RoCE NETWORKING - Remote DMA over Converged Ethernet

Wire speed, low latency over common 40GbE media allows better than native SAS disk I/O performance via IBM General Parallel File System (renamed Spectrum Scale) developed for IBM High Performance Computing platforms like BlueGene and Watson. Each ESS I/O Power8 node has two Mellanox 40GbE ConnectX3 RoCE adapters and three 6Gb SAS adapters to dual split back-plane IBM DSC7000 disk arrays. Solution has more than enough bandwidth to support better than native SAS performance to four concurrent Power8 analytic client workstations.

IBM ESS - Elastic Storage Server

ESS solution uses one Power8 controller and two Power8 I/O nodes networked via dual parallel Mellanox 40GbE networks (TCP/IP and RoCE). Each Power8 I/O node has three 6Gb LSI SAS adapters with dual paths to two split back-plane IBM DSC7000 JBOD arrays which is able to source and sink GPFS data at high network utilization rates. Each JBOD has 58 4Tb HDD and a 500Gb SSD for meta-data. The Spectrum Scale RAID software provides redundancy and best-in-class performance compared to other networked storage technologies (like Fibre Channel SAN and CEPH).

IBM POWER8 HMC LPAR - Bare Metal Logical PARtition

Each IBM Power8 S822A workstation is symmetrically configured with dual NUMA processors and a split internal SAS storage back-plane with dual RAID controllers and dual two-port Mellanox 40GbE RoCE network adapters. An IBM Power HMC (Hardware Maintenance Console) is used to Logically PARtition (LPAR) each physical Power8 system in two identical logical parts without the use of a hypervisor and any associated performance penalties. Almost a two for the price of one cost advantage.