IMPALA – An agile and advanced data analytics platform

Ram Pisipati | 281.483.5786 | ram.pisipati@nasa.gov
Data Integration Analytics & Analytics Director
Human Health & Performance Contract

TCC Data Analytics Workshop
October 3, 2017
NASA Human Health and Performance Overview

Goal: Enable Successful Space Exploration by Minimizing the Risks of Spaceflight Hazards
HH&P System Evolution & Current Challenges

Crew Health & Safety Program

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury Gemini Apollo</td>
<td>Skylab</td>
<td>Belmont Report</td>
<td>Space Shuttle</td>
<td>Common Rule</td>
<td>ISS</td>
<td>Electronic Medical Record</td>
</tr>
</tbody>
</table>

Data Archiving |

Current Challenges

Data captured in old formats are not query-able by new tools

Legacy systems were designed to capture & store information, not for analytics

Information Silos - Data spread across multiple systems of record with limited access

Big gap between “useful” data vs. “usable” data

Health Exposures

Pre-NASA Exposures - ASCAN Training - Active Astronaut - Retirement

Terrestrial Exposures
What is IMPALA?

IMPALA is a big data analytics platform that provides data cleansing, enrichment & cataloging tools with an intuitive, interactive and visual user interface to quickly & easily acquire, organize, analyze & decide using data in the reservoir, then search/find, share, collaborate and publish analysis results.

- Advanced analytics (e.g., trending, forecasting, natural language processing)
- Business intelligence (e.g., interactive scorecards, dashboards, reports)
- Data mining, exploration & discovery
- Refined & reusable data sets (e.g., cardio, exercise, VIIP...)

Data Analytics, Scalable Platform, Skills + HH&P Data & Domain Expertise → Insights & Mission Effectiveness

IMPALA - Information Management Platform for anALytics & Aggregation
Connecting the Consumer with the Capability

Application User: Which reports, analysis and what decisions use my application data?

IT Developer: What is the downstream impact of changing this data table?

HH&P Systems: Raw data

Data Ingestion & Processing

Data Refinery

Data Reservoir

Reporting & Analysis

Actionable Events

Actionable Information

Actionable Insights

Data Governance & Common Data Dictionary

Catalog & Index Data

HH&P Data Lab

Share Discovery Outputs

Data Steward: Can I trust that the data is representative of source data and what was done to it?

Data Scientist: I want to discover the correlation between space flight and headaches. What data do I have?

Data Analyst: How was the BP measurement taken? Where did this CO₂ data come from?

Administrator: Does this system comply with NASA’s data management policies?
Core Capabilities

- High Performing Hadoop Data Reservoir
- Rules based Data Preparation & Transformation
- Fine grained security & access control

- Data Tagging & Meta data
- Data Catalog
- Secure Collaboration Workspace

- Self Service Data Lab
- Advanced Analytics & Modeling
- Data Exploration, Mining & Visualization

Enabling HH&P Wide Analytics & Insights Through Secure Collaboration

National Aeronautics and Space Administration
# IMPALA – Periodic Table of Capabilities

<table>
<thead>
<tr>
<th>Capability</th>
<th>Description</th>
<th>Icon</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Ingestion</strong></td>
<td>Capability to connect and import data at scale from a variety of data sources and formats for immediate use or long term storage. Data can be streamed in real time or ingested in batch.</td>
<td><img src="https://example.com" alt="DS" /></td>
<td><img src="https://example.com" alt="Ds" /> Data in SQL DB</td>
</tr>
<tr>
<td><strong>Data Catalog</strong></td>
<td>Self service central location for metadata that is searchable, discoverable and editable. The data catalog connects the physical data to the HH&amp;P business terms &amp; glossary.</td>
<td><img src="https://example.com" alt="Dn" /></td>
<td><img src="https://example.com" alt="Dn" /> Data in NoSQL DB</td>
</tr>
<tr>
<td><strong>Data wrangling</strong></td>
<td>The process of transforming, cleaning and getting data into a usable format.</td>
<td><img src="https://example.com" alt="Me" /></td>
<td><img src="https://example.com" alt="Me" /> Merge &amp; Join data</td>
</tr>
<tr>
<td><strong>Data Analytics</strong></td>
<td>Platform to create, collaborate and perform advanced analytics on data sets across many data domains.</td>
<td><img src="https://example.com" alt="Cd" /></td>
<td><img src="https://example.com" alt="Cd" /> Create Datasets</td>
</tr>
<tr>
<td><strong>Data visualization</strong></td>
<td>A set of tools to visualize and analyze data at scale.</td>
<td><img src="https://example.com" alt="E" /></td>
<td><img src="https://example.com" alt="E" /> Visual Data Exploration</td>
</tr>
<tr>
<td><strong>Data security</strong></td>
<td>Built in security functions in IMPALA to secure, protect and preserve data.</td>
<td><img src="https://example.com" alt="MI" /></td>
<td><img src="https://example.com" alt="MI" /> Machine Learning</td>
</tr>
<tr>
<td><strong>Data Governance</strong></td>
<td>Control and authority (planning, monitoring &amp; enforcement) over the management of IMPALA’s data reservoir content.</td>
<td><img src="https://example.com" alt="La" /></td>
<td><img src="https://example.com" alt="La" /> Link Analysis</td>
</tr>
</tbody>
</table>

**Developer**

- Data Scientist
- Data Steward
- Data Analyst
- Administrator
- Developer
- Governance Board
Lessons Learned

➢ Build the data platform to adapt, not to last
  • Big Data field is new and technologies are evolving

➢ Don’t underestimate data integration challenges
  • The longer it takes to find data, the less valuable it becomes

➢ Design for security & privacy from get go
  • Focus on security, privacy and data management

➢ Develop talent organically
  • Insights require deep understanding of data