MR Research on the Cloud – A Flywheel/Columbia University Case Study

Can Akgun, PhD[1] and John Thomas Vaughan, PhD[2]
[1]Flywheel Exchange, LLC
[2]Columbia University MR Research Center

Introduction/Motivation

- MR researchers are challenged with managing large data volumes, computationally intensive analyses, and the need to share this data through collaboration inter or intra-institutional [1].
- A robust and scalable computational environment is necessary to effectively manage current and previously acquired data, automate tasks, and scale processing to meet today's researchers' needs.
- **Columbia University's Zuckerman Mind, Brain, Behavior Institute (ZI) [2]** has tackled these challenges by partnering with **Flywheel[3]** to automatically capture all MR data and store it (with tertiary data) in the Google Cloud Platform (GCP)[4] to take advantage of long-term data archiving and scalable, on-demand computation.
- ZI is the first institute within **Columbia University's MR Center (CMRRC),** a network of institutions that includes Columbia's Irving Medical Center, School of Engineering and Applied Sciences, the Nathan Kline Institute for Psychiatric Research, and the New York State Psychiatric Institute.

System Architecture Overview

Core: Backend server that provides all core functionality, such as storing files, maintaining a database, and managing permissions and security.

Data Connectors: Background processes that automate data workflows by monitoring device APIs or file systems for new data

- Automated capture of data from MR and other devices
- De-identification of data to meet privacy requirements
- File encryption of data during storage and transfer

Data Management: Management of acquired data and metadata, including labeling for ML workflow and elastic search

Compute Engine: Queuing and managing processing jobs for data analysis and data converters, known as "Gears."

Data Access and Analysis: Open API, web application, library of software development kits (SDKs), and command-line-interface (CLI)

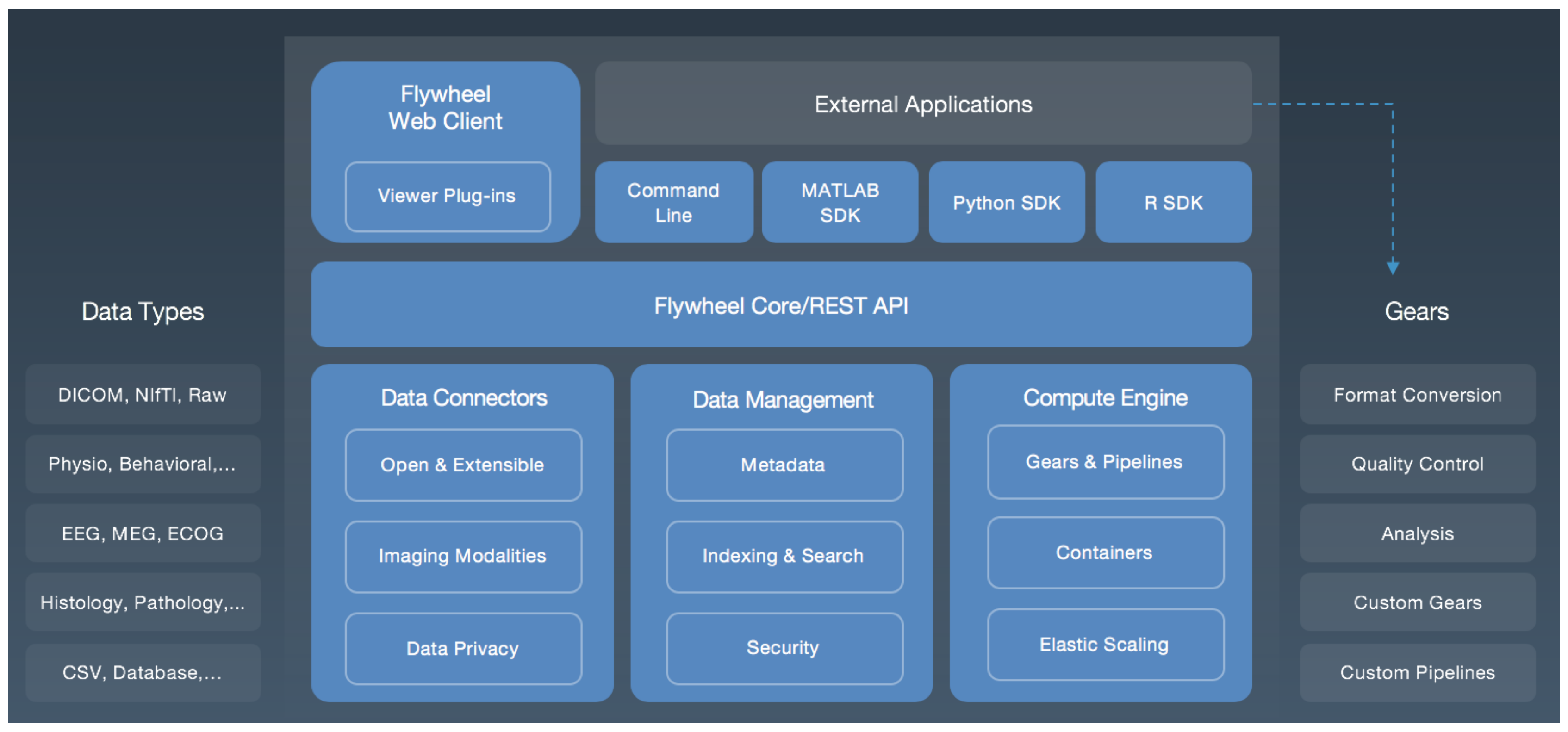

Figure 1. Architecture of Flywheel software platform

References

[1] B. Wandell, et. al, Data management to support reproducible research. **ARXIV, Quantitative Biology** – Quantitative Methods, Bibliographic Code: 2015arXiv150206900W, [2] https://zuckermaninstitute.columbia.edu/, [3] https://flywheel.io/, [4] https://cloud.google.com

Columbia University's MR Research on the Cloud

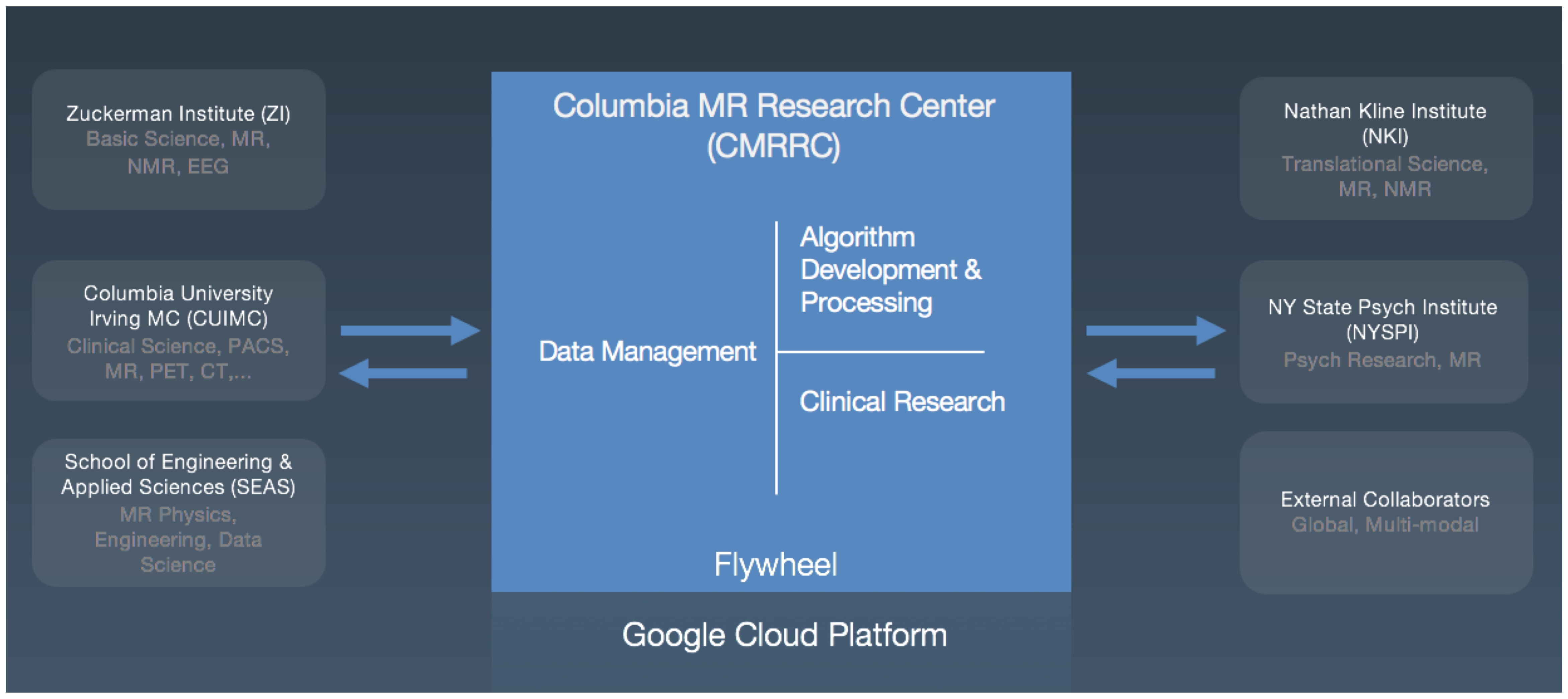
Zuckerman Institute (ZI)
Basic Science, MR, NMR, EEG
Columbia University Irving MC (CUIMC)
Clinical Science, PACS, MR, PET, CT,...
School of Engineering & Applied Sciences (SEAS)
MR Physics, Engineering, Data Science
Columbia MR Research Center (CMRRC)
Data Management
Algorithm Development & Processing
Clinical Research
Flywheel
Google Cloud Platform
Nathan Kline Institute (NKI)
Translational Science, MR, NMR
NY State Psych Institute (NYSPI)
Psych Research, MR
External Collaborators
Global, Multi-modal

Figure 2. Research components at Columbia's MR Research Center

Zuckerman Institute's MR Core Fully Integrated in the Cloud

Data Capture

- Automated data ingestion and routing to GCP from two 3T MR scanners since June 1, 2017
- PET/MR scanner, Bruker animal scanner, EEG system, Behavioral data
- Expanded to 16 MR scanners at CMRRC by 2020

Curation

- Research hierarchy and workflow for each investigator's Lab
- Metadata capture, AI/Machine learning labeling
- Indexing & Elastic search, data sub-grouping ("collections") for ML training sets
- Quality controls, data viewing and downloading

Computation

- Automated pre-processing (data conversion, classification) and Quality Assurance (QA) gears (SNR, spike plots, motion) per project
- Full pipeline processing with common open-source gears (FSL, FreeSurfer, HCP, etc.) with scalable virtual machine (VM) deployment on GCP
- Custom algorithms (versioned and provided supporting meta-data such as author, maintainer, and description)

Collaboration

- Secure sharing of data with internal and external collaborators
- Access controls
- BIDS data support and per project templating

Zuckerman Statistics as of 10/2018

Number of scanners	3
Number of labs	34
Number of unique projects	130
Number of sessions	2000+
Number of gears run	112,000+
External institutions collaborating with ZI	10+
Number of gears available on platform	30+
Number of concurrent jobs that can be run in parallel on GCP	50

Conclusions

- The acquisition, management, sharing, and analysis schema described herein meet the requirement of Columbia University's scientific data management network.
- By leveraging the cloud, researchers now have access to scalable computation and long-term archiving.
- Data, algorithms and analysis at the CMRRC are shared through a distributed network to support and promote collaboration.