



The Analytics Pipeline and Data Flow

September 20, 2018

Linton Ward, PhD
IBM Distinguished Engineer
OpenPower Cognitive Solutions

Emmanuel Macron Talks to WIRED About France's AI Strategy

[Nicholas Thompson](#) [business](#) 03.31.18 06:00 am



EM: I think artificial intelligence will disrupt all the different business models and it's the next disruption to come. So I want to be part of it. Otherwise I will just be subjected to this disruption without creating jobs in this country. So that's where we are. And there is a huge acceleration and as always the winner takes all in this field.

<https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/>

New DHS S&T Program Targets Internet, Critical Infrastructure Disruption

The new program – the Predict, Assess Risk, Identify (and Mitigate) Disruptive Internet-scale Network Events (PARIDINE) project – aims to study Network/Internet-scale Disruptive Events (NIDE), which can cut internet or network connectivity, leading to disruptions of “energy and water systems, the finance sector, commerce, and public safety and emergency communications systems, as well as other essential systems.”

<https://www.meritalk.com/articles/new-dhs-st-program-paridine/>

State Department Looking for Platform to Track, Analyze Online Info

The State Department has issued a request for information for systems that collect relevant online information to “analyze and track global developments in (near) real-time.” ... The State Department listed its needs for a monitoring system, including: aiding the ability to verify the credibility of a source; ensuring the accuracy of machine-generated content from different languages; and distributing information quickly.

<https://www.meritalk.com/articles/state-department-looking-for-platform-to-track-analyze-online-info/>

GAO Releases Updated Cyber Risk Report

The Government Accountability Office (GAO) today released an updated version of a report it issued in July detailing major cybersecurity challenges facing the Federal government and critical actions needed to address them.

<https://www.meritalk.com/articles/gao-releases-updated-cyber-risk-report/>

House Bill to Codify CDM Moves to Senate

“Cyberattacks are escalating at an alarming rate, making it vital that our Federal agencies have access to programs and tools to help mitigate these risks,”

<https://www.meritalk.com/articles/house-bill-to-codify-cdm-moves-to-senate/>

NIST Wants to Know: Can You Trust Your IoT?

The draft publication outlines 17 trust-related issues “that may negatively impact the adoption of IoT products and services,” spanning scalability, predictability, difficult in measurement, lack of certification criteria, all the way down to usability, performance, and reliability.

DHS CIO Says Priorities Include Modernization, Workforce, Supply Chain

The Department of Homeland Security (DHS) is focused on modernizing its mindset to tackle a host of pressing issues including reducing its reliance on legacy systems, competing to attract cybersecurity talent, and combating supply chain threats, said DHS CIO John Zangardi today at the Billington Cybersecurity Summit.

“We’re in a very, very different world than we have been in the past,” said Zangardi.

“I’ve been in government for a long time. We’re really good at routine. But cyber threats are asymmetrical. The adversary’s not thinking about routine, the adversary is thinking about how to do things differently.”

<https://www.meritalk.com/articles/dhs-cio-says-priorities-include-modernization-workforce-supply-chain/>

HHS CTO Report Calls Data Silos to Task

A new [report](#) from the Department of Health and Human Services’ (HHS) CTO calls out the department and its individual agencies for keeping their data in silos, and calls for a department-wide data governance framework.

“Whether surveillance, survey, or claims data, HHS expends an enormous amount of financial resources to report on the state of the health of the population it serves,”

<https://www.meritalk.com/articles/hhs-cto-report-calls-data-silos-to-task/>

SEC Looking for Social Media Monitoring Tool

The Securities and Exchange Commission (SEC) on Thursday issued a [solicitation](#) for “a web-based subscription to a Commercial-Off the Shelf (COTS) social media monitoring tool that provides emailed alerts to SEC staff based on keyword searches for relevant topics with ability to monitor social media sites.”

<https://www.meritalk.com/articles/sec-looking-for-social-media-monitoring-tool/>



PRESIDENT'S MANAGEMENT AGENDA

The Administration is developing a Federal Data Strategy to leverage data as a strategic asset to grow the economy, increase the effectiveness of the Federal Government, facilitate oversight, and promote transparency.



Strategy 1: Enterprise Data Governance. Set priorities for managing Government data as a strategic asset, including establishing data policies, specifying roles and responsibilities for data privacy, security, and confidentiality protection, and monitoring compliance with standards and policies ...

•Strategy 2: Access, Use, and Augmentation.

Develop policies and procedures and incentive investments that enable stakeholders to effectively and efficiently access and use data assets by: (1) improving dissemination, making data available more quickly and in more useful formats; (2) maximizing the amount of non-sensitive data shared with the public; and (3) leveraging new technologies and best practices to increase access to sensitive or restricted data while protecting the privacy, security, and confidentiality, and interests of data providers.

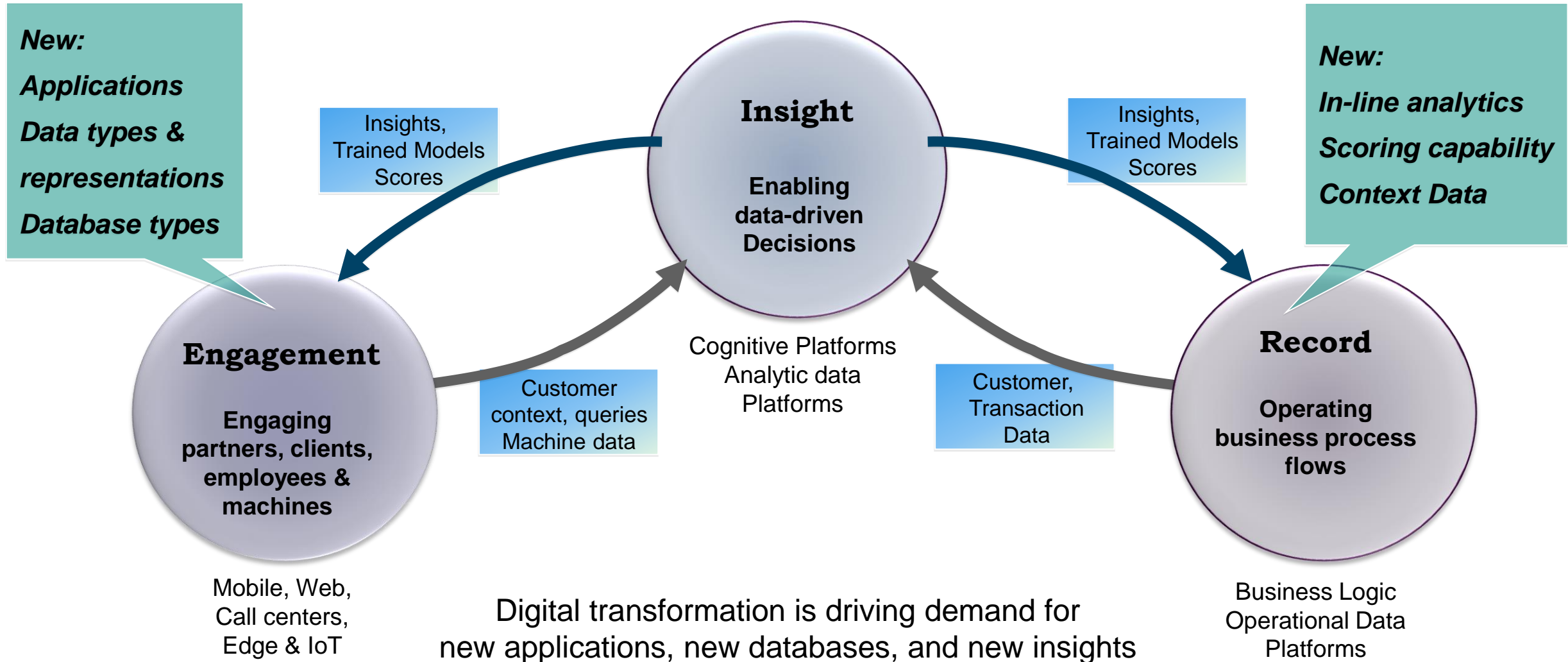
•Strategy 3: Decision-Making and Accountability.

Improve the use of data assets for decision-making and accountability for the Federal Government, including both internal and external uses. This includes: (1) providing high quality and timely information to inform evidence-based decision-making and learning; (2) facilitating external research on the effectiveness of Government programs and policies which will inform future policymaking; and (3) fostering public accountability and transparency

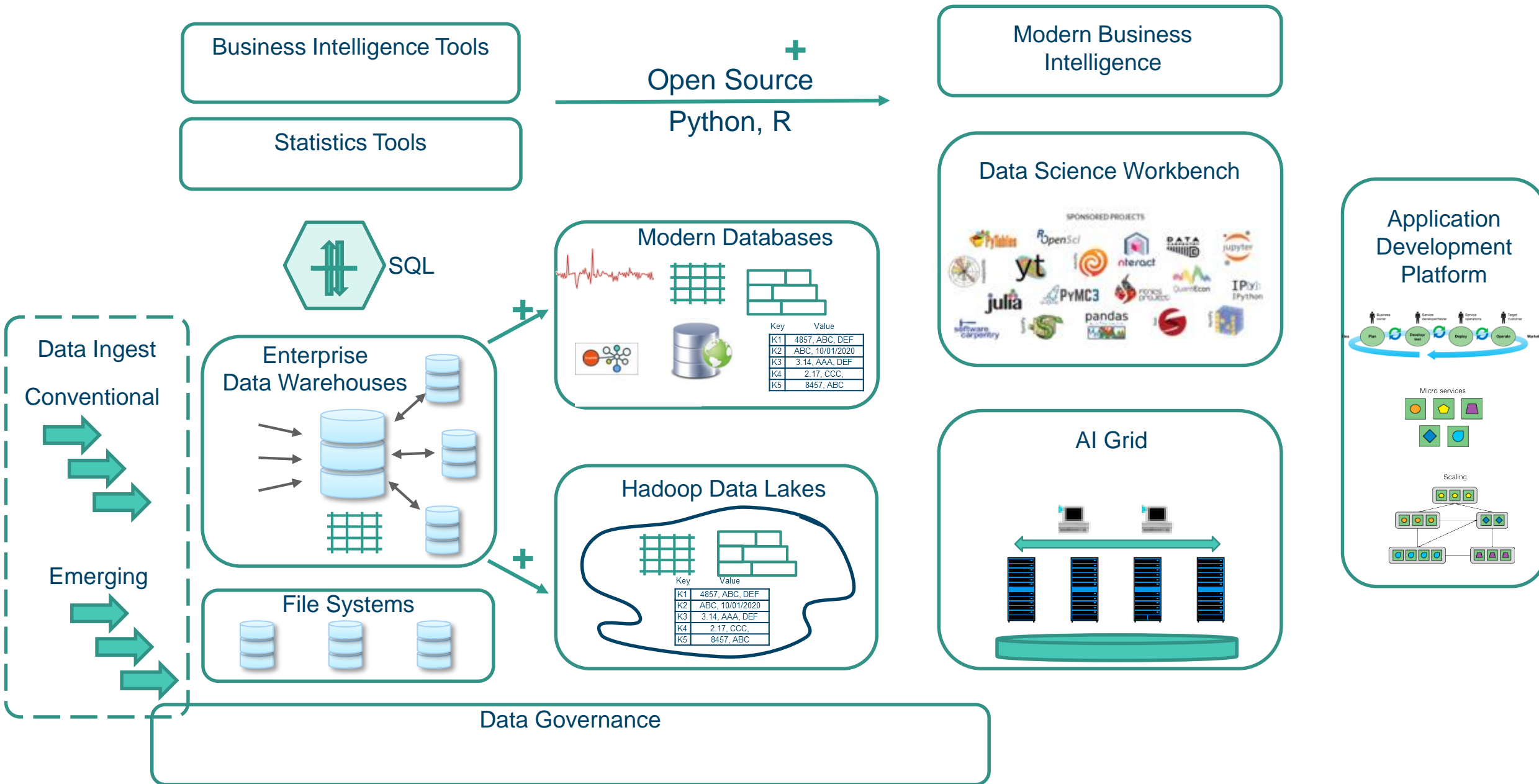
Strategy 4: Commercialization, Innovation, and Public Use.

Facilitate the use of Federal Government data assets by external stakeholders at the forefront of making Government data accessible and useful through commercial ventures, innovation, or for other public uses. This includes use by the private sector and scientific and research communities; by states, localities, and tribes for public policy purposes; for education; and in enabling civic engagement.

Application Transformation



Systems of Insight Landscape



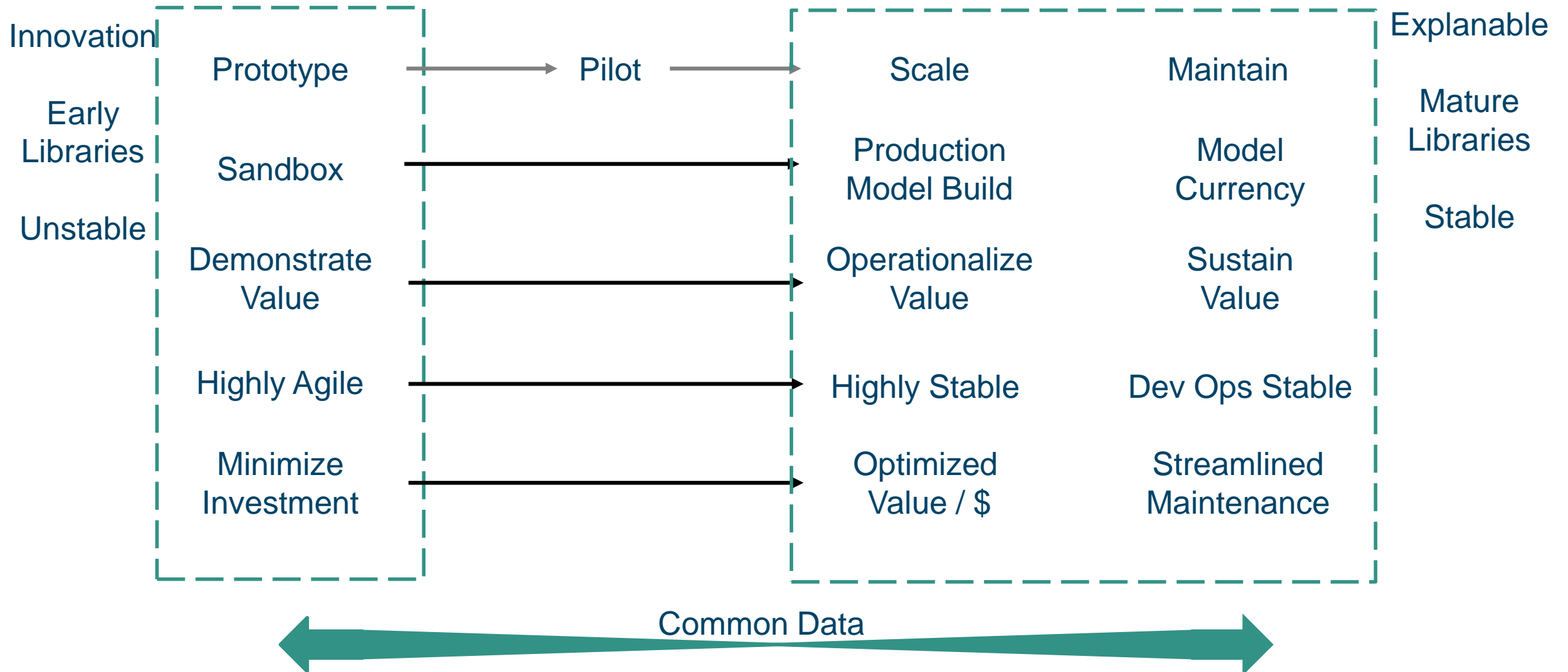
Success with analytics projects (ways to succeed)

How do we derisk analytics projects?

Clarity on the question	Apply critical thinking techniques with buy-in
Enable faster exploration	The data science workbench: create an <i>ad hoc</i> workflow quickly
Enable quicker win	Data science sandbox: prototype from data scientist rather than presentation alone
Scale to production	AI Grid: multi-tenant, high stability, high efficiency cluster

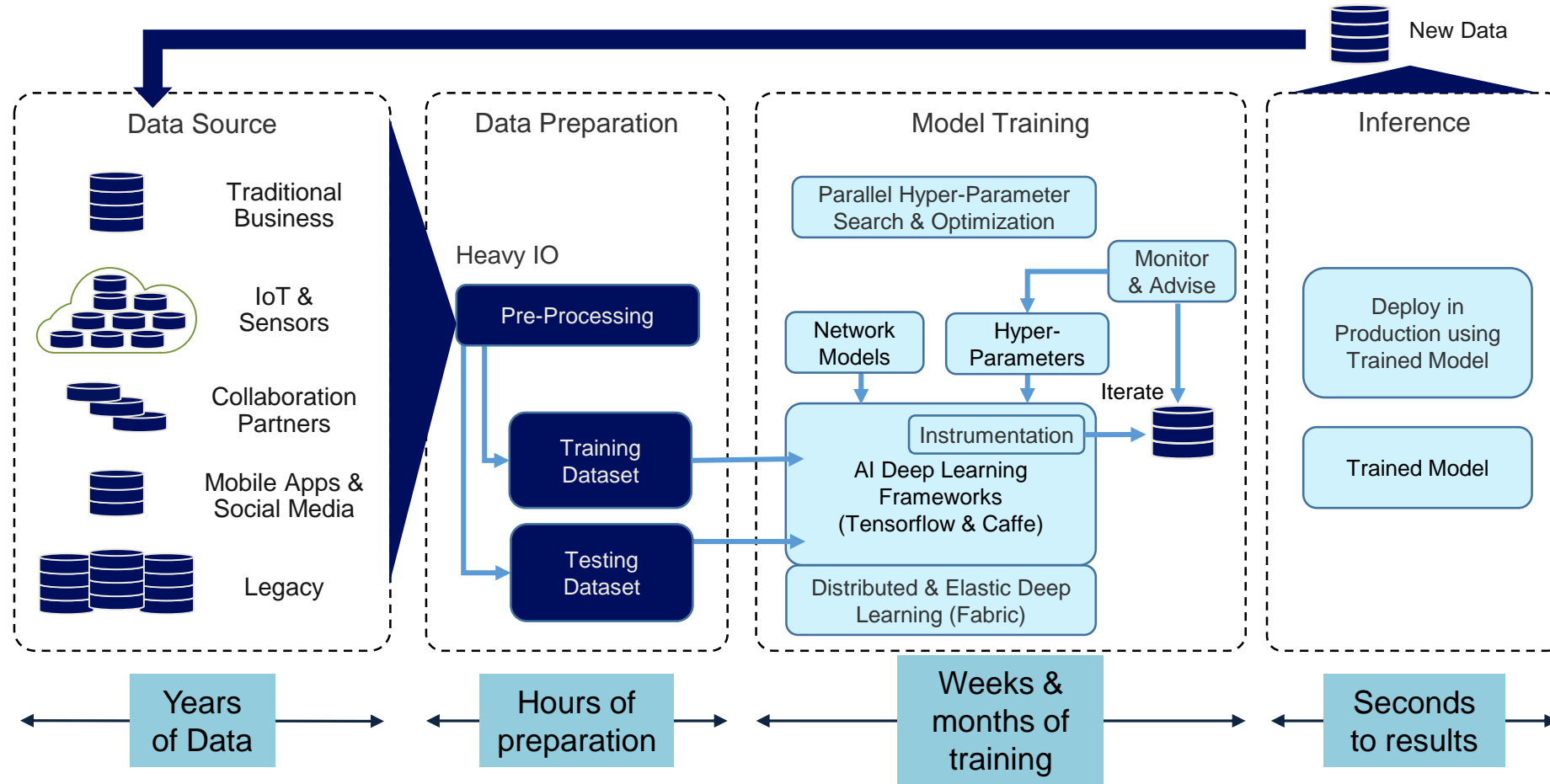
Cognitive Platform: Analytic Project Lifecycle

Progression from Data Science Workbench to operationalized insights



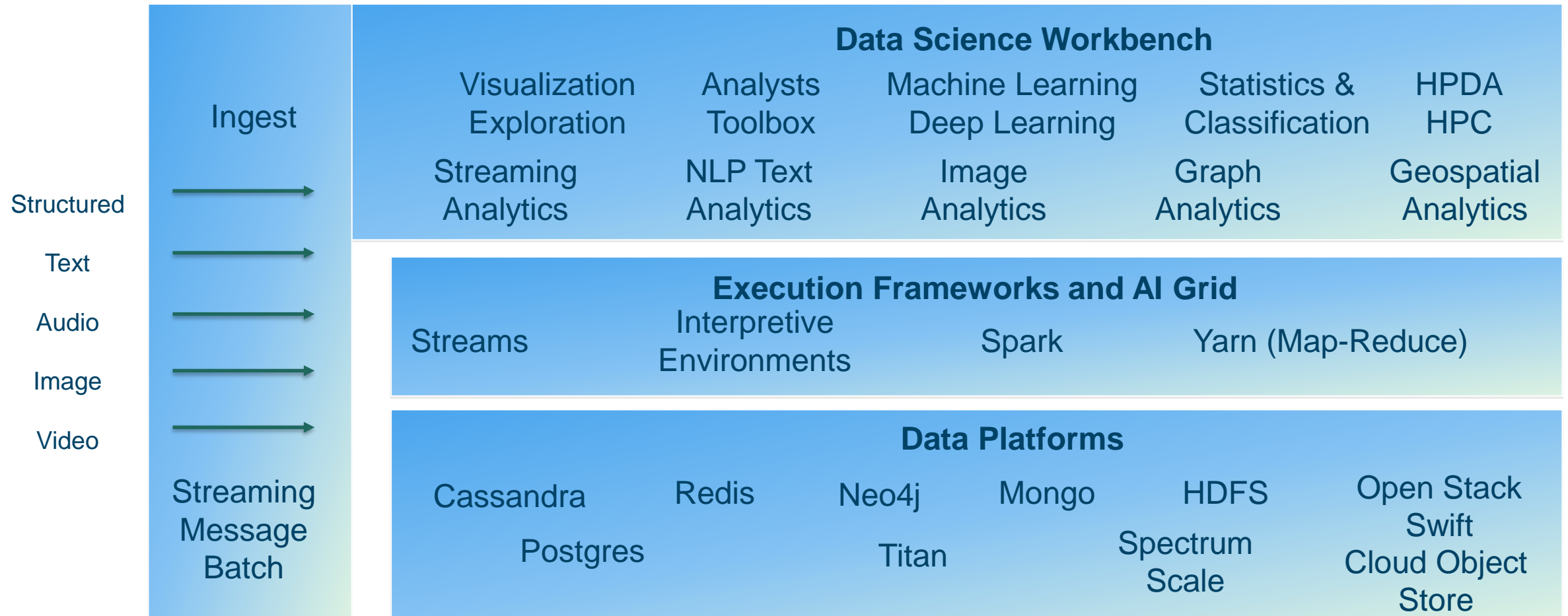
The Data Science Workbench

Workload flow and data flow are key to results

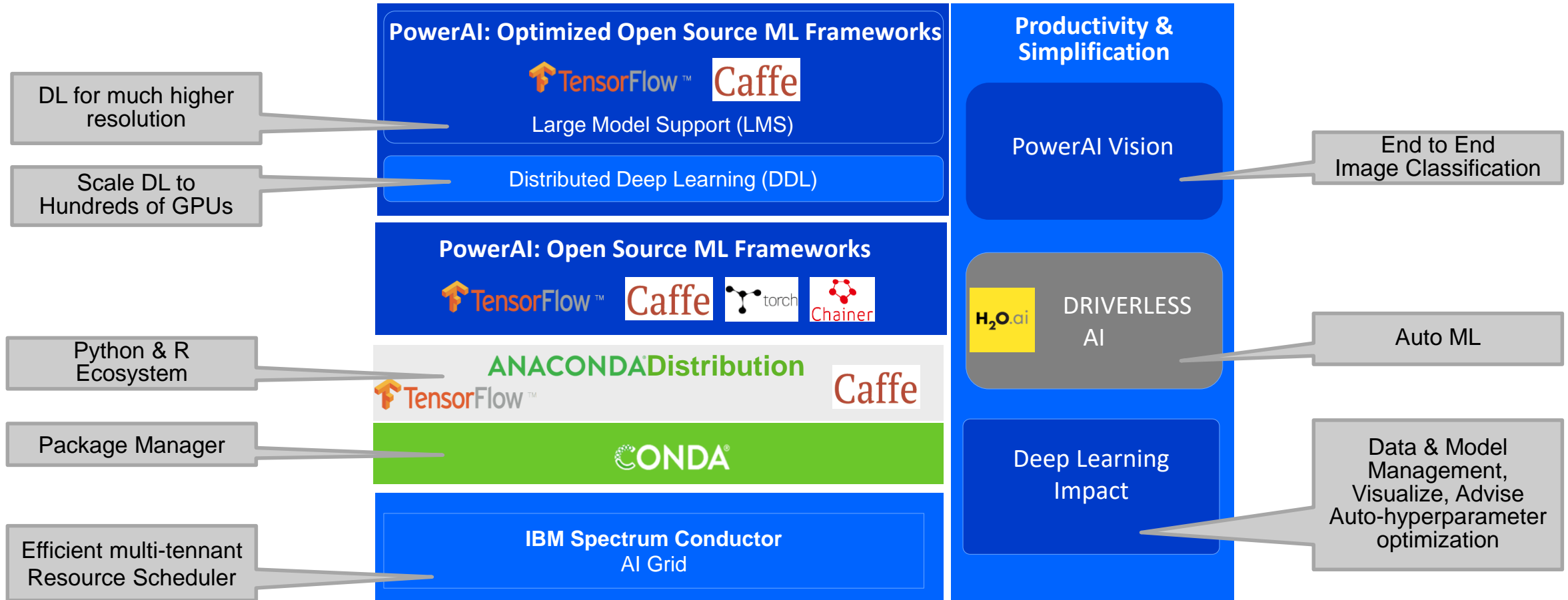


Cognitive Systems – Capabilities in the Data Science Workbench

The Data Science Workbench comprises a set of capabilities



PowerAI Enterprise



ANACONDA Accelerates Adoption of Open Data Science for Enterprises

ANACONDA[®]

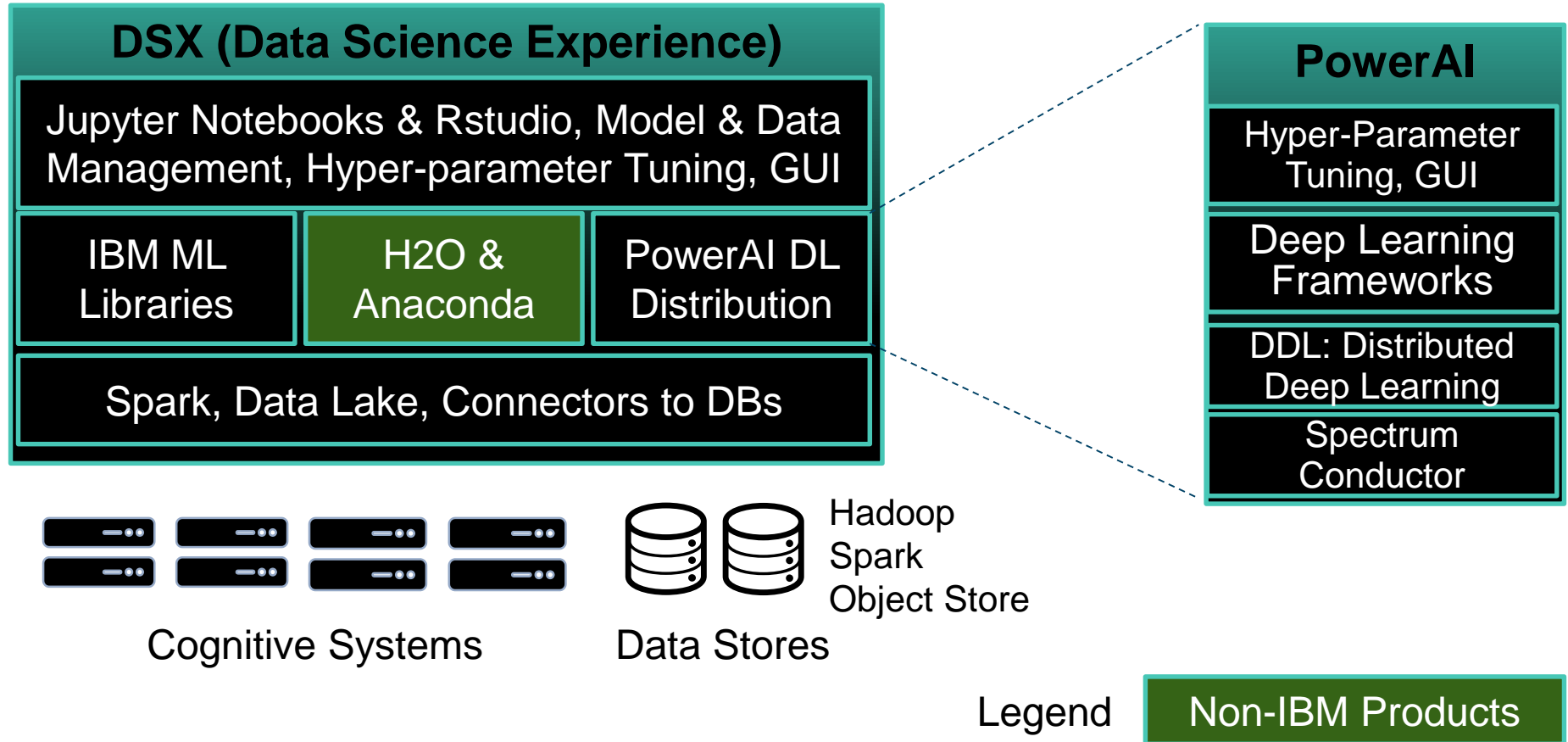
PYTHON & R OPEN SOURCE ANALYTICS

NumPy	SciPy	Pandas	Scikit-learn	Jupyter/IPython	
Numba	Matplotlib	Spyder	Numexpr	Cython	Theano
Scikit-image	NLTK	NetworkX	IRKernel	dplyr	shiny
ggplot2	tidyr	caret	PySpark	& 720+ packages	



- Easy to install
- Agile data exploration
- Powerful data analysis
- Simple to collaborate
- Accessible to everyone

IBM AI / Data Science Workbench: DSX Local



The AI Grid

IBM Software Defined Infrastructure

Multi-scale Infrastructure for High Performance Computing & Analytics

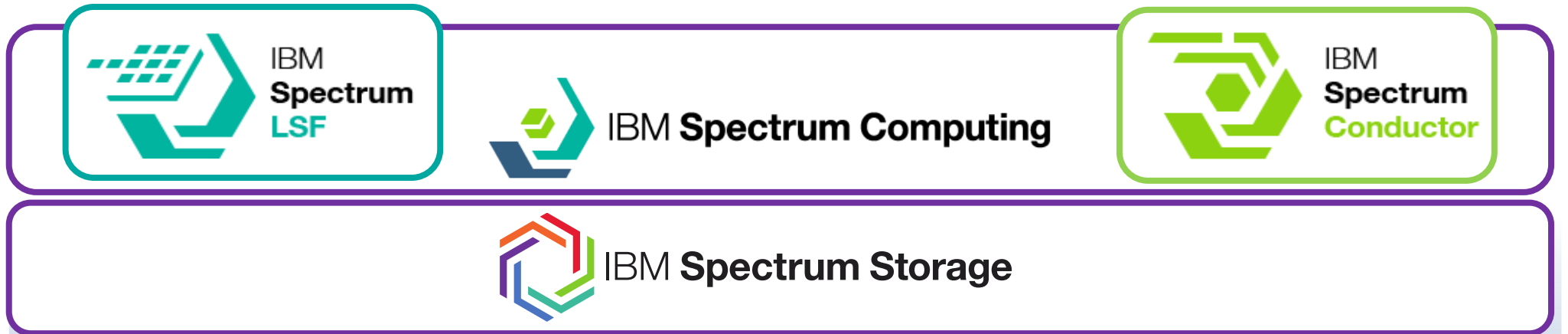
High Performance Computing
Design / Simulation / Modeling

'New-gen Workloads'
Hadoop, Spark, Containers

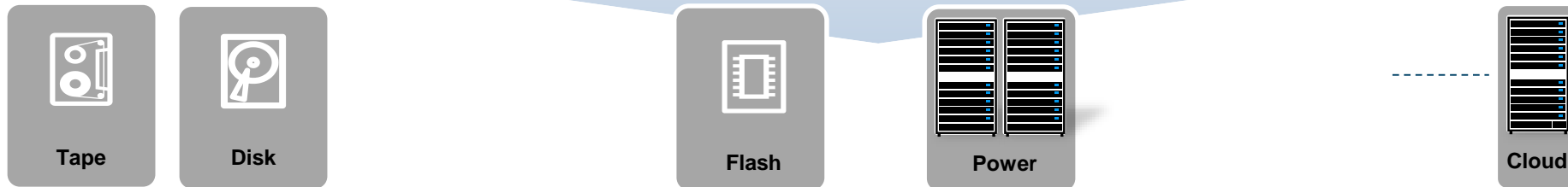
**Workload
Aware
Scheduling**

**Shared
Resource
Management**

**Shared
Multi-tier
Data Management**



Servers & Storage



Hybrid Cloud Infrastructure

IBM Spectrum Conductor

Secure Multi-tenant, deploy and manage modern computing frameworks & services

Faster Time to Results

- Proven High-performance scalable resource and job scheduler
- Multitenant resource sharing

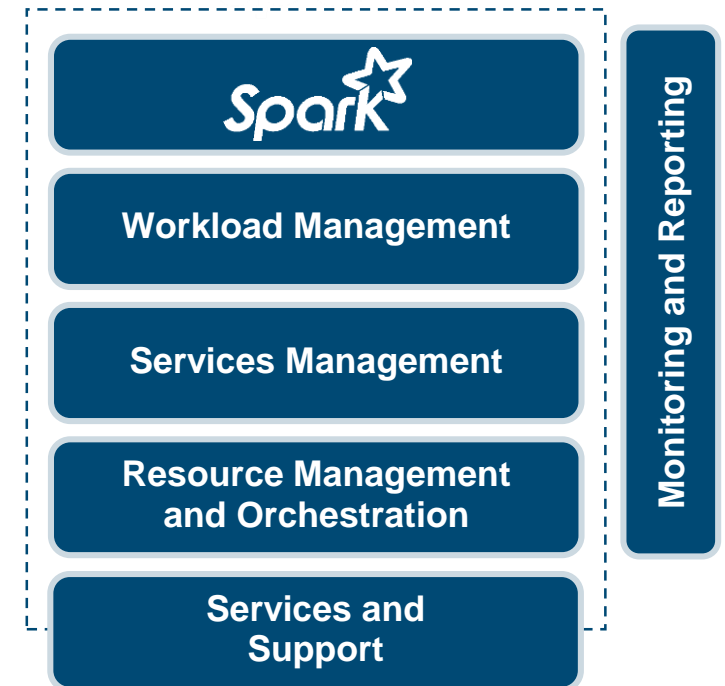
Simplified Deployment & Management

- Complete solution: scheduling, monitoring, alerting, reporting & diagnostics
- Lifecycle management supporting multiple concurrent and different versions

Lower Infrastructure Costs with Optimized Resource Sharing

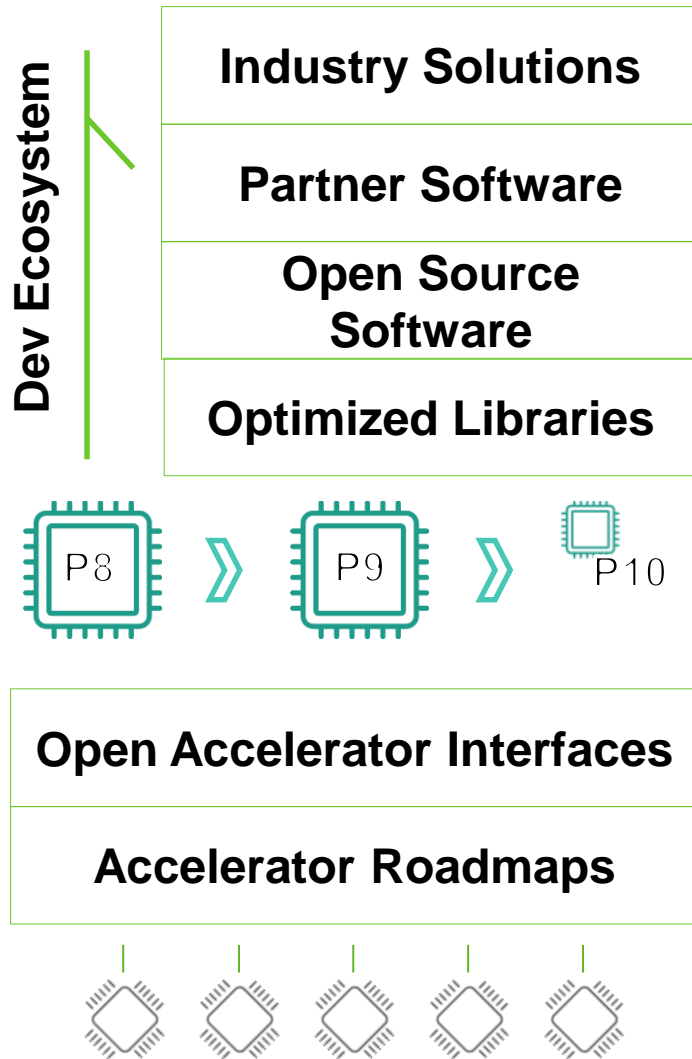
Coming Soon

- Enhanced Notebook & Anaconda Integrations
- Job Dependencies
- DSX Integration
- Fine Grained Resource Allocation



Delivering Value for Data Science

Cognitive Systems are built with optimized hardware and software

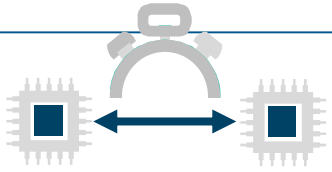


Not Just About Hardware Design

It's about co-optimized



which ***just work*** for Machine Learning, Deep Learning, and AI



Faster Training Time with Distributed Deep Learning

Shape Boundary
Attenuation
Recognition

9 Days

Shape Boundary
Attenuation
Recognition

4 Hours
4 Hours
4 Hours
4 Hours
4 Hours
4 Hours
4 Hours
4 Hours

What will you do?
Iterate more and create more accurate models?
Create more models?
Both?

4 Hours
4 Hours
4 Hours
4 Hours

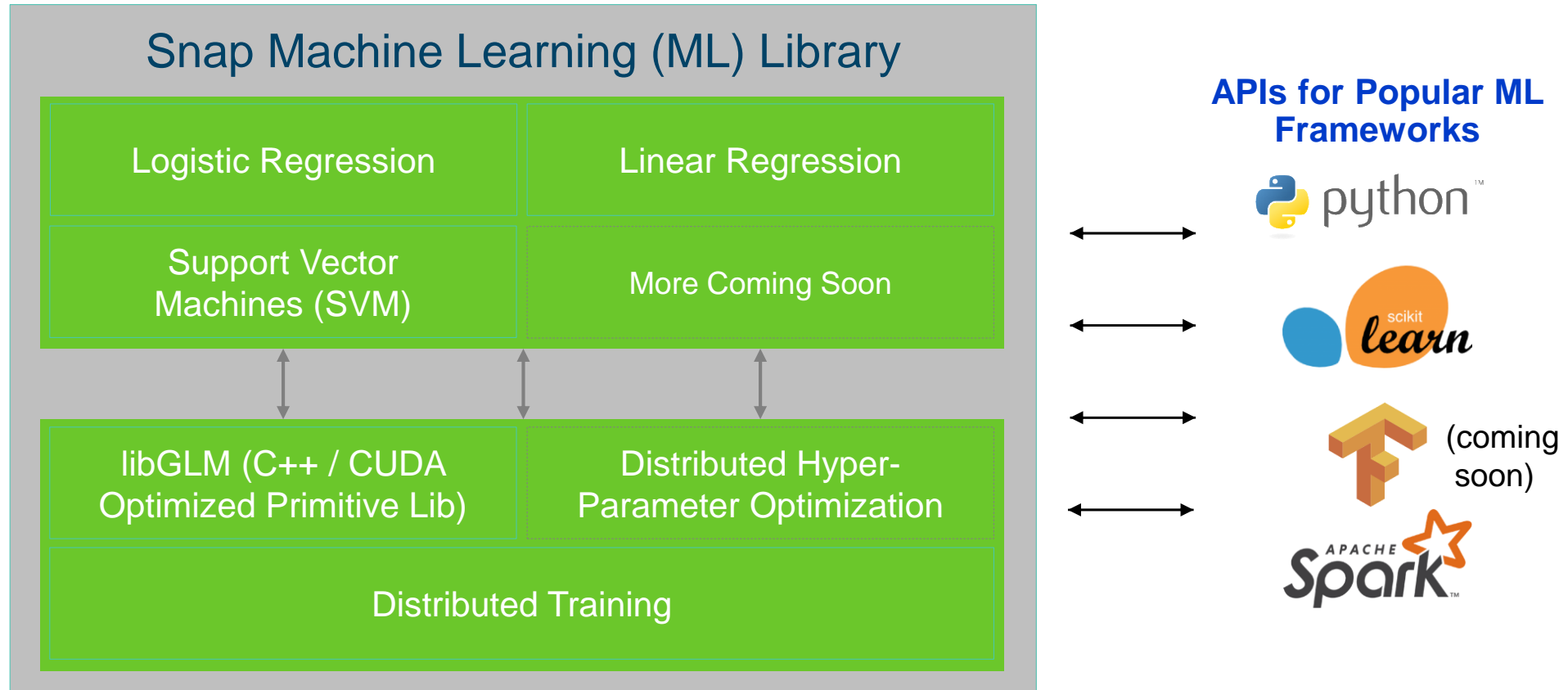
100x

Learning runs with Power 9*

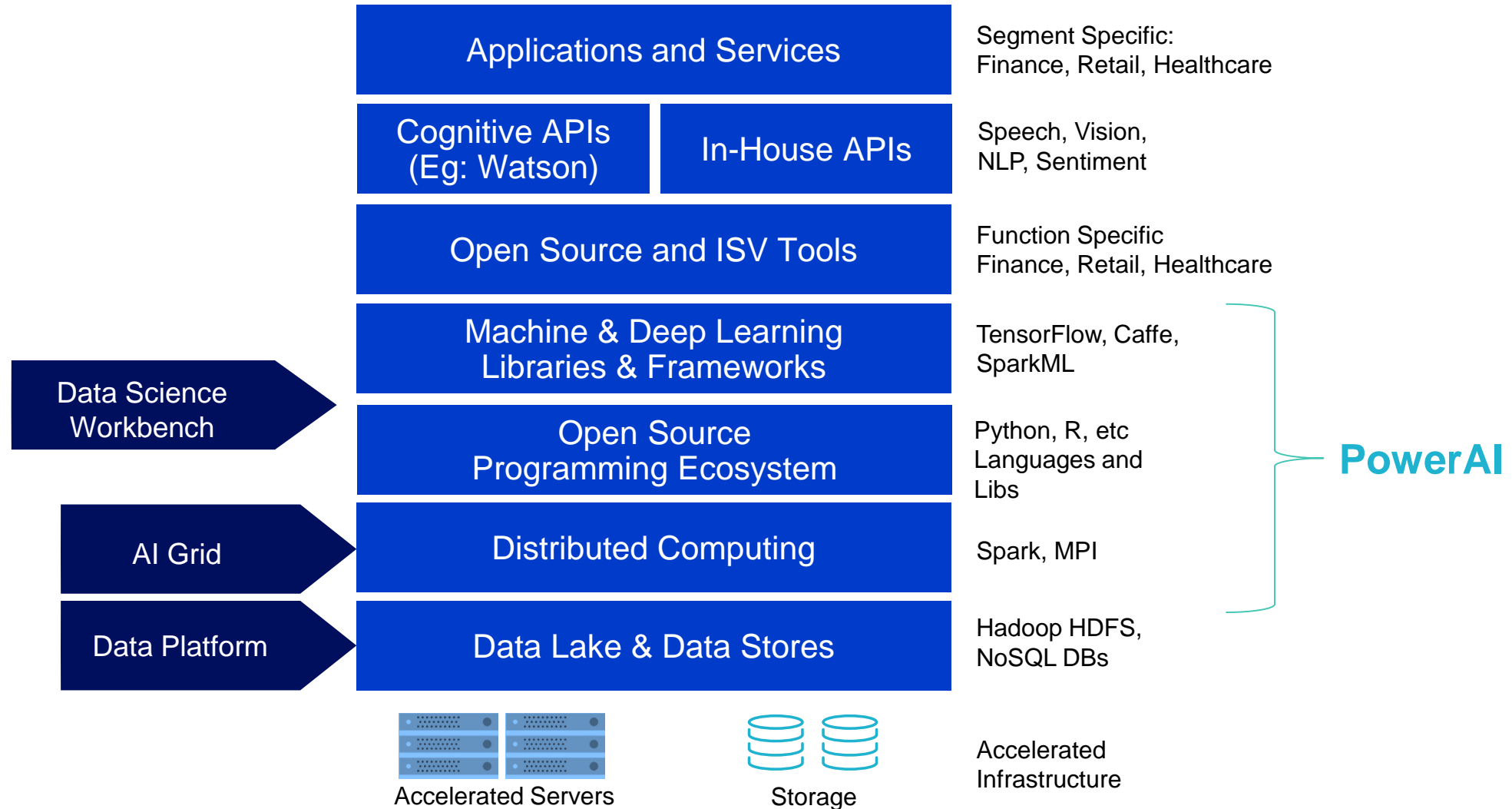
54x

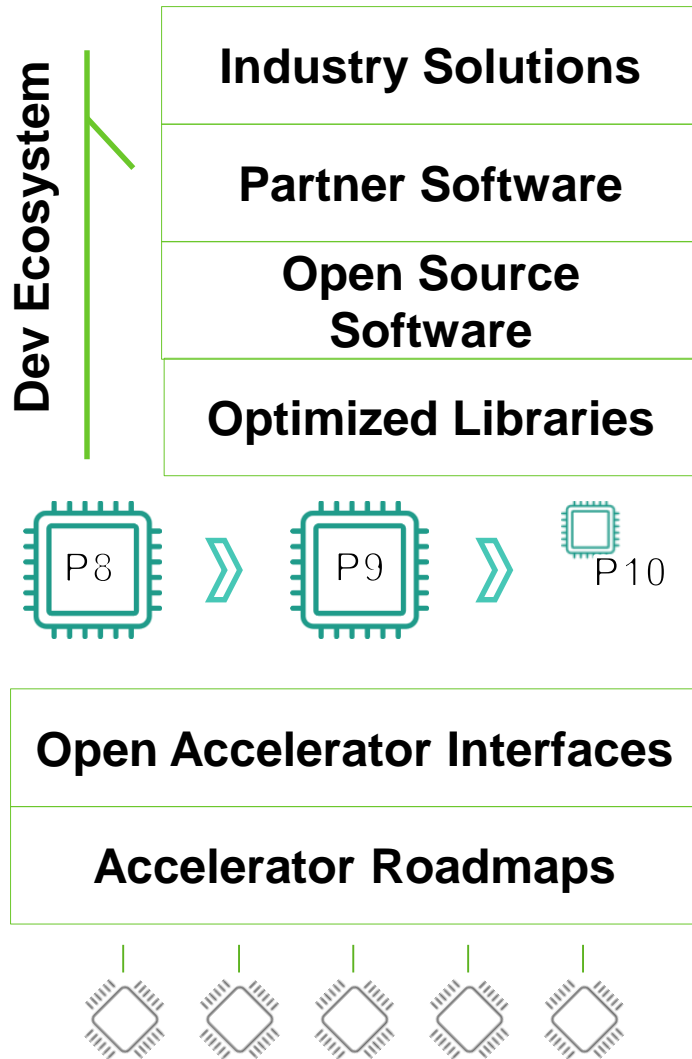
Learning runs with Power 8

Distributed GPU-Accelerated Machine Learning Library



An Optimized AI Infrastructure Stack





Time to value for new intelligence

Data Science Productivity

Data Productivity

AI for the rest of us

Solve larger problems

Solve previously intractable problems

“We can do new science”