Spectrum Scale and Compute

Data Architecture for Big Data, Analytics, and Cognitive Clouds

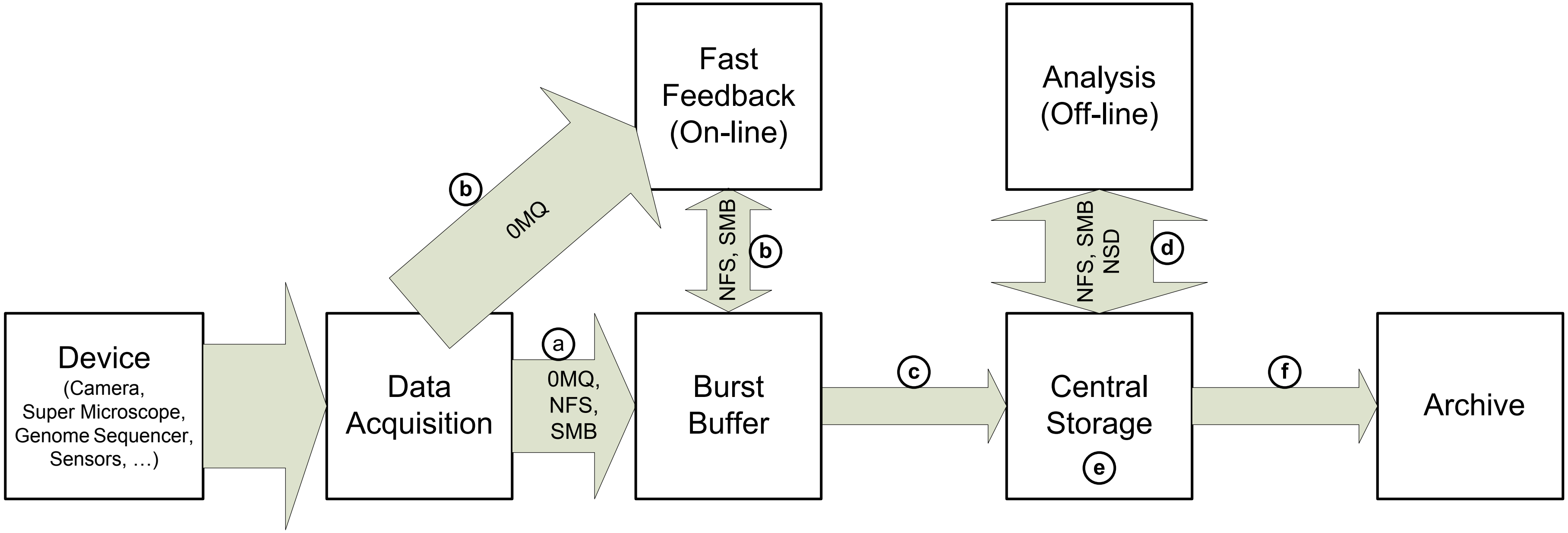
Agenda

- Data Centric Use Cases
- Data Intensive Science Challenges
- AI Hierarchy of Needs
- Data Architecture

Spectrum Scale
Data Intensive Use Cases

- Instrument Driven Science and Healthcare
  - ALS and CryoEM
  - Radio Astronomy
  - Weather
  - Genomics
- Data Driven Engineering
  - Modeling and simulation results (i.e. visualization)
  - Sensor data analysis
  - Financial − market data analysis and reporting
  - Supply chain efficiency
- Big Data
  - Market Insights
  - Operational efficiency (including IT)
- Cognitive
  - Personalized Medicine
  - Autonomic Driving
  - Natural Language Processing

Typical workflow



a) Real-time data ingest (data acquisition)

b) Visualization and near real-time analysis (online processing)

c) Data movement from Burst Buffer to Central Storage

d) Deep analysis (offline processing)

e) Data management of Central Storage

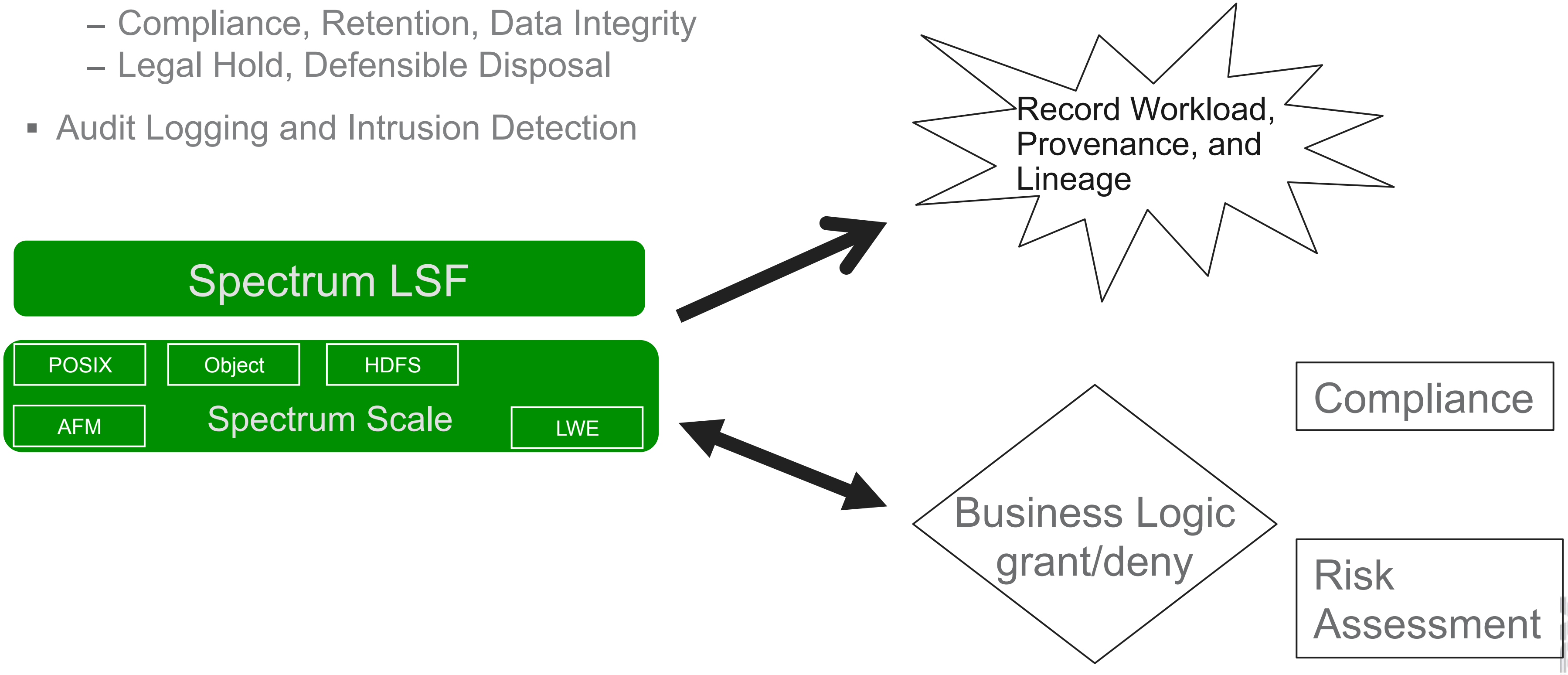
f) Long-term data archiving

Note: User/Scientists need access to data during each stage of the workflow.

Note: The workflow distills data to the relevant insight by increasing the ratio of content/data (=Pipeline of forgetting the unimportant).

Data and Workflow Challenges

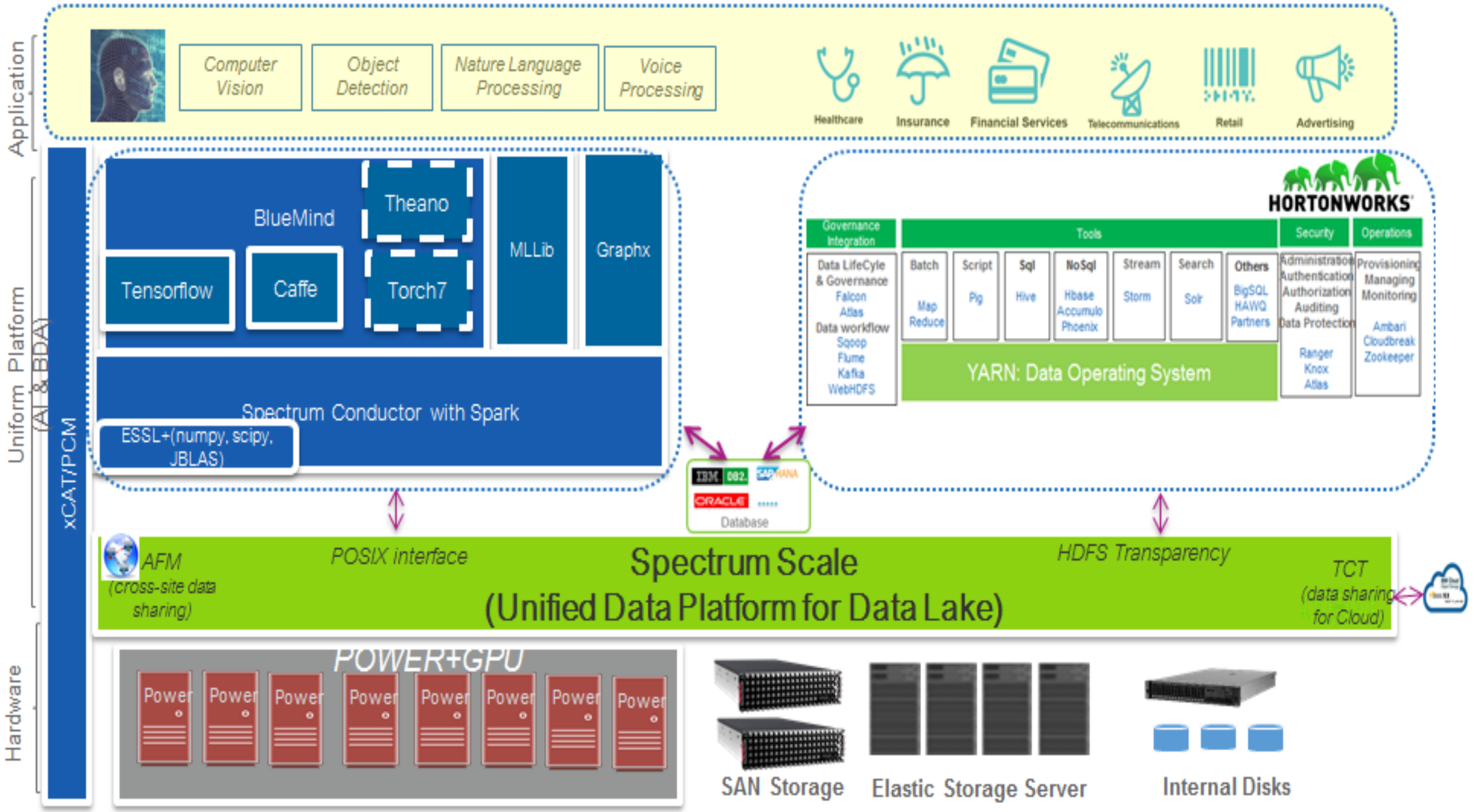
- Provenance
  - Reproducibility
    - Scientific Workflows: transformations, interpretations, analyses
  - Lineage
    - Origin
    - Ownership
    - Usage
- Governance
  - Compliance, Retention, Data Integrity
  - Legal Hold, Defensible Disposal
- Audit Logging and Intrusion Detection



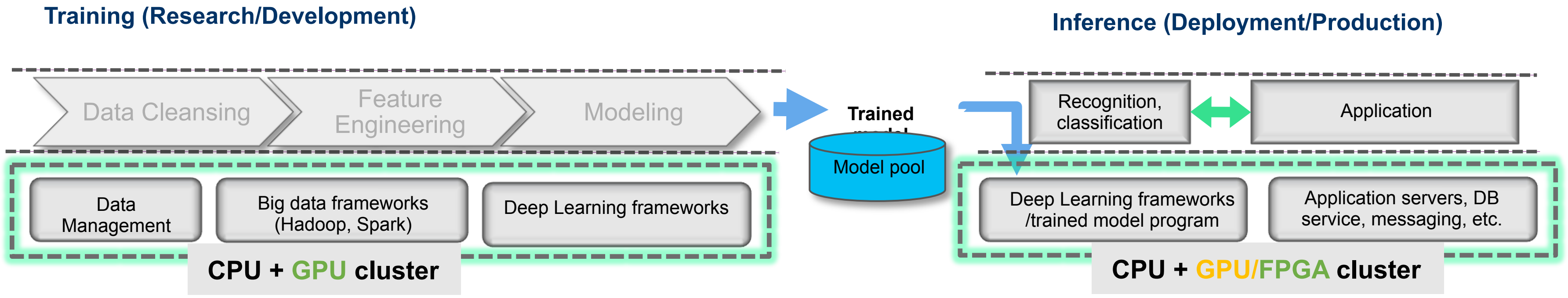
Spectrum Scale Solution for Cognitive and BDA (Best Practices)

Solution Key Values:

- Support long-term rapid increasing big data with extreme scaling for file system
- Fast analytics results from in-place analytics without data movement
- Easy maintenance from centralized storage management for multiple Hadoop cluster
- Support internal disk based for entry level customer(less than 100TB data size) and scale to PB level in ESS



Cognitive Workloads with Spectrum Scale Data Ocean



Training (Research/Development)

Data Cleansing → Feature Engineering → Modeling

Data Management | Big data frameworks (Hadoop, Spark) | Deep Learning frameworks

CPU + GPU cluster

Trained model

Model pool

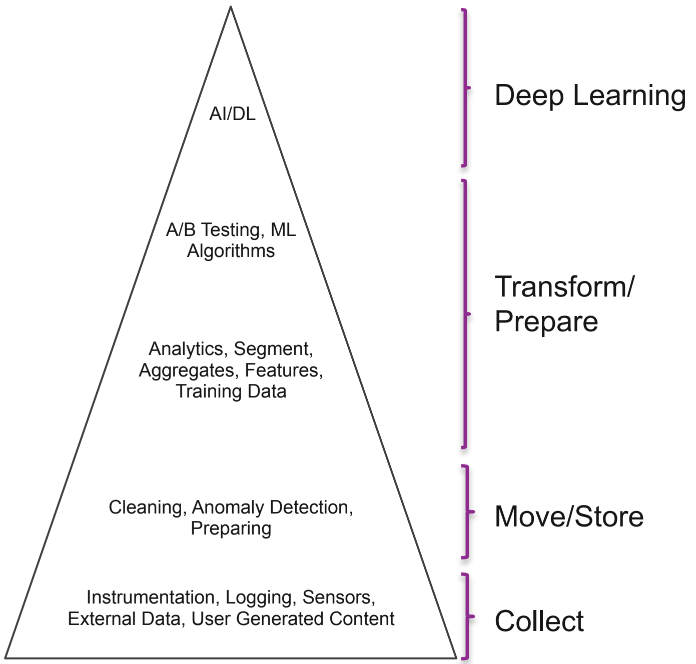
Inference (Deployment/Production)

Recognition, classification ↔ Application

Deep Learning frameworks /trained model program | Application servers, DB service, messaging, etc.

CPU + GPU/FPGA cluster

Data Hierarchy in Deep Learning



AI/DL

A/B Testing, ML Algorithms

Analytics, Segment, Aggregates, Features, Training Data

Cleaning, Anomaly Detection, Preparing

Instrumentation, Logging, Sensors, External Data, User Generated Content

Deep Learning

Transform/ Prepare

Move/Store

Collect

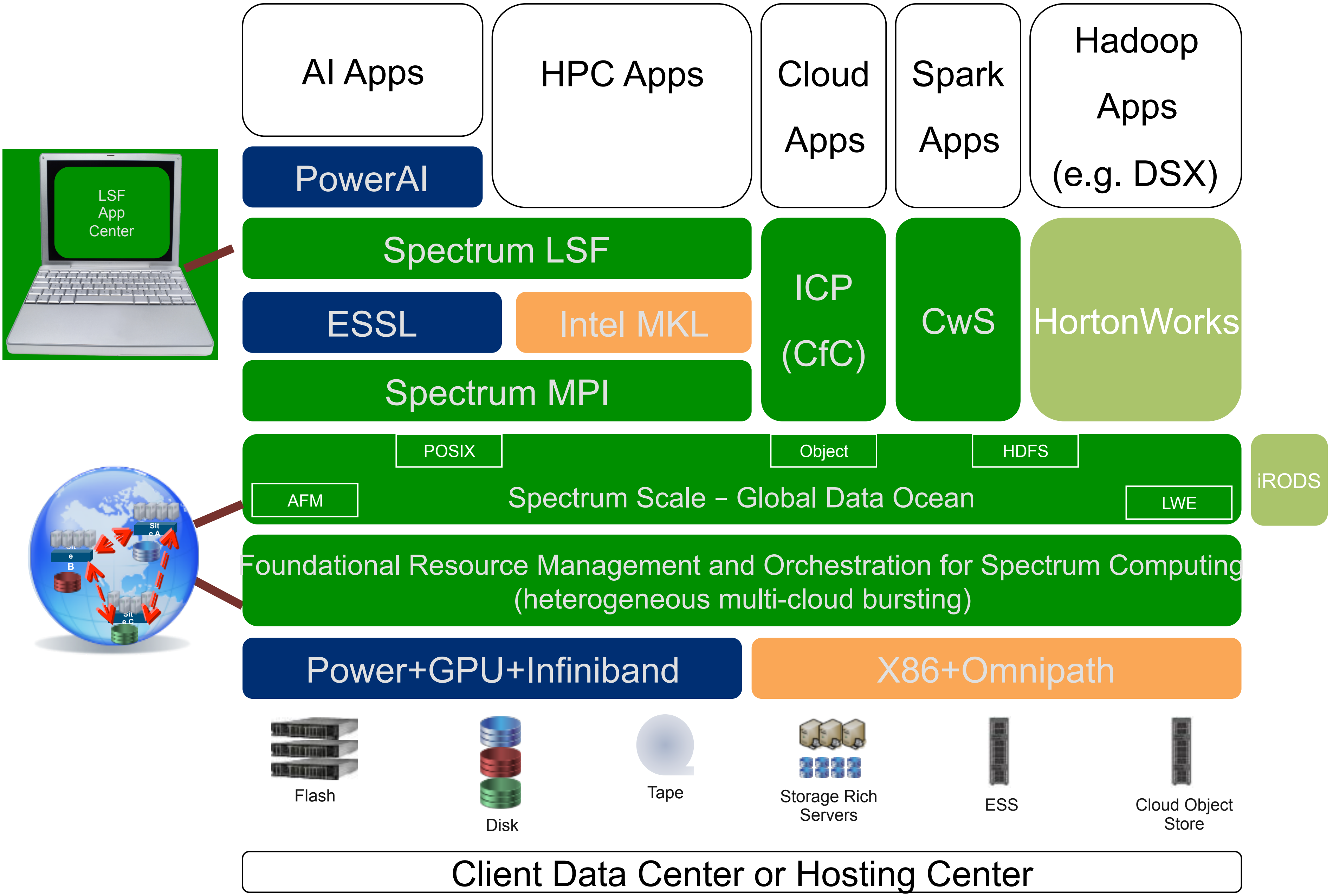
Phase	Tools and Applications
Deep Learning	*Frameworks*: TensorFlow(Apache), Caffe(BSD), Torch(BSD), Theano(free), CNTK(free), Neon(Apache) *IDE*: ***IBM Power AI Enterprise/CwS with deep learning**, Nvidia Digits, ***Watson** *Spectrum Scale*: POSIX for Power AI; need to evaluate performance
Transform/Prepare	*Machine Learning*: IBM SPSS, ***IBM DSX**, SAS *Spectrum Scale*: SPSS works over GPFS POSIX/HDFS; SAS works over GPFS POSIX
Move/Store	*Platform*: Hadoop, Spark, POSIX/NFS/SMB *ETL*: ***Talend** *Spectrum Scale*: POSIX/HDFS interface; need to evaluate performance
Collect	Variable data ingestion end devices *Spectrum Scale*: Supports Swift/S3 object interface for data ingest

***need go to market partnerships**

IBM Confidential

Ref: Monica Rogati – AI Hierarchy of Needs

Cloud Software Architecture



AI Apps

HPC Apps

Cloud Apps

Spark Apps

Hadoop Apps (e.g. DSX)

PowerAI

Spectrum LSF

ESSL

Intel MKL

Spectrum MPI

ICP (CfC)

CwS

HortonWorks

POSIX

Object

HDFS

AFM

Spectrum Scale – Global Data Ocean

LWE

iRODS

Foundational Resource Management and Orchestration for Spectrum Computing (heterogeneous multi-cloud bursting)

Power+GPU+Infiniband

X86+Omnipath

Flash

Disk

Tape

Storage Rich Servers

ESS

Cloud Object Store

Client Data Center or Hosting Center