Science-Aware Dynamic Data Delivery at the Exascale

Torre Wenaus
Physics Applications Software Group
Brookhaven National Laboratory

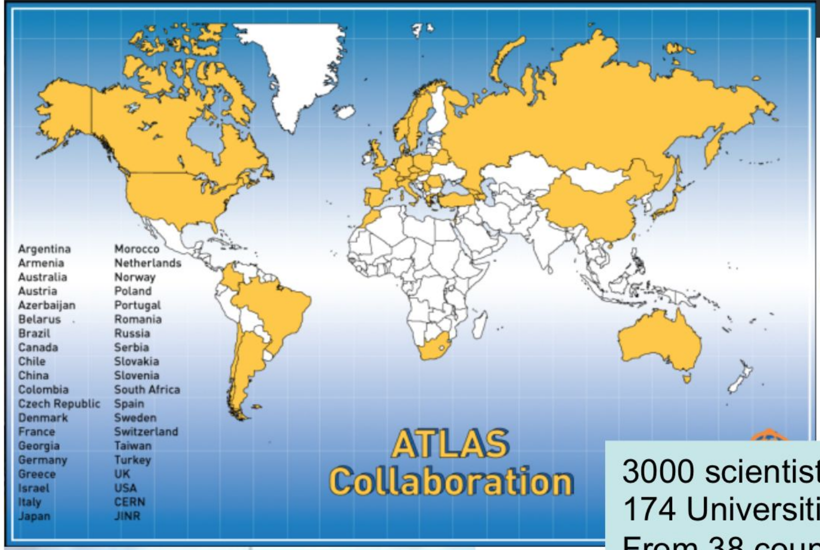
STREAM 2015 Workshop
IUPUI, Indianapolis
October 28, 2015

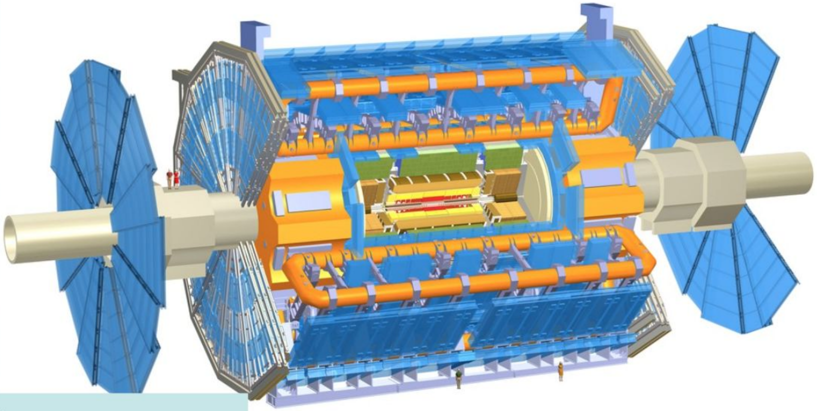
BROOKHAVEN
NATIONAL LABORATORY

a passion for discovery

Office of Science
U.S. DEPARTMENT OF ENERGY

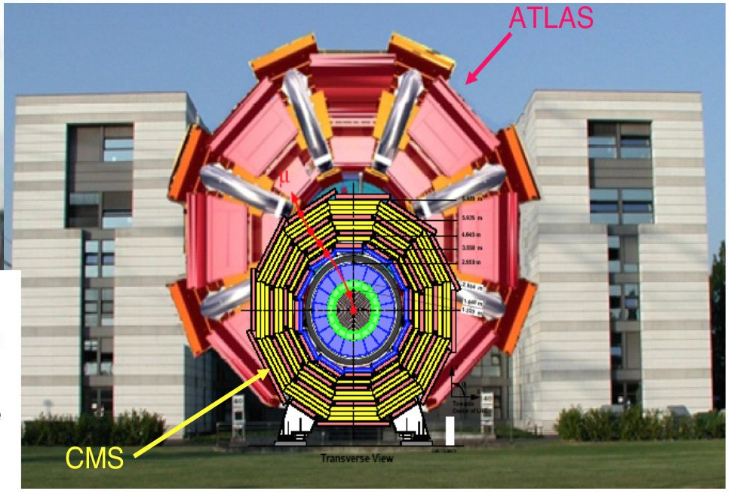
The ATLAS Experiment at the LHC

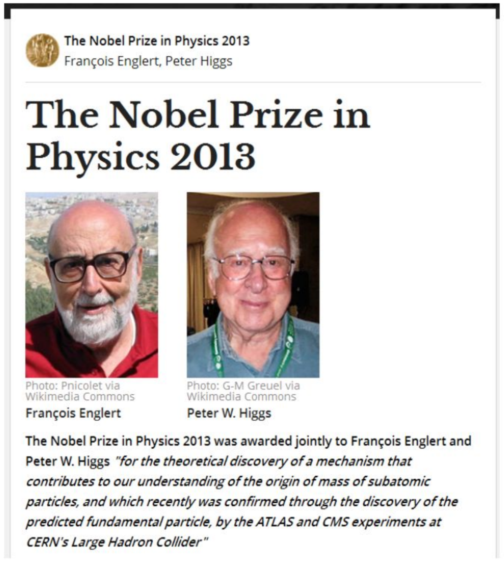




3000 scientists
174 Universities and Labs
From 38 countries
More than 1200 students

- **ATLAS has 44 meters long and 25 meters in diameter, weighs about 7,000 tons. It is about half as big as the Notre Dame Cathedral in Paris and weighs the same as the Eiffel Tower or a hundred 747 jets**





The Nobel Prize in Physics 2013
François Englert, Peter Higgs

The Nobel Prize in Physics 2013

Photo: Pnicolet via Wikimedia Commons
François Englert

Photo: G-M Greuel via Wikimedia Commons
Peter W. Higgs

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

Large Scale Data Intensive Processing: The LHC Data Torrent

Drop of water: Roughly 0.1 mL

New physics rate ~ 0.00001 Hz

Event Selection :

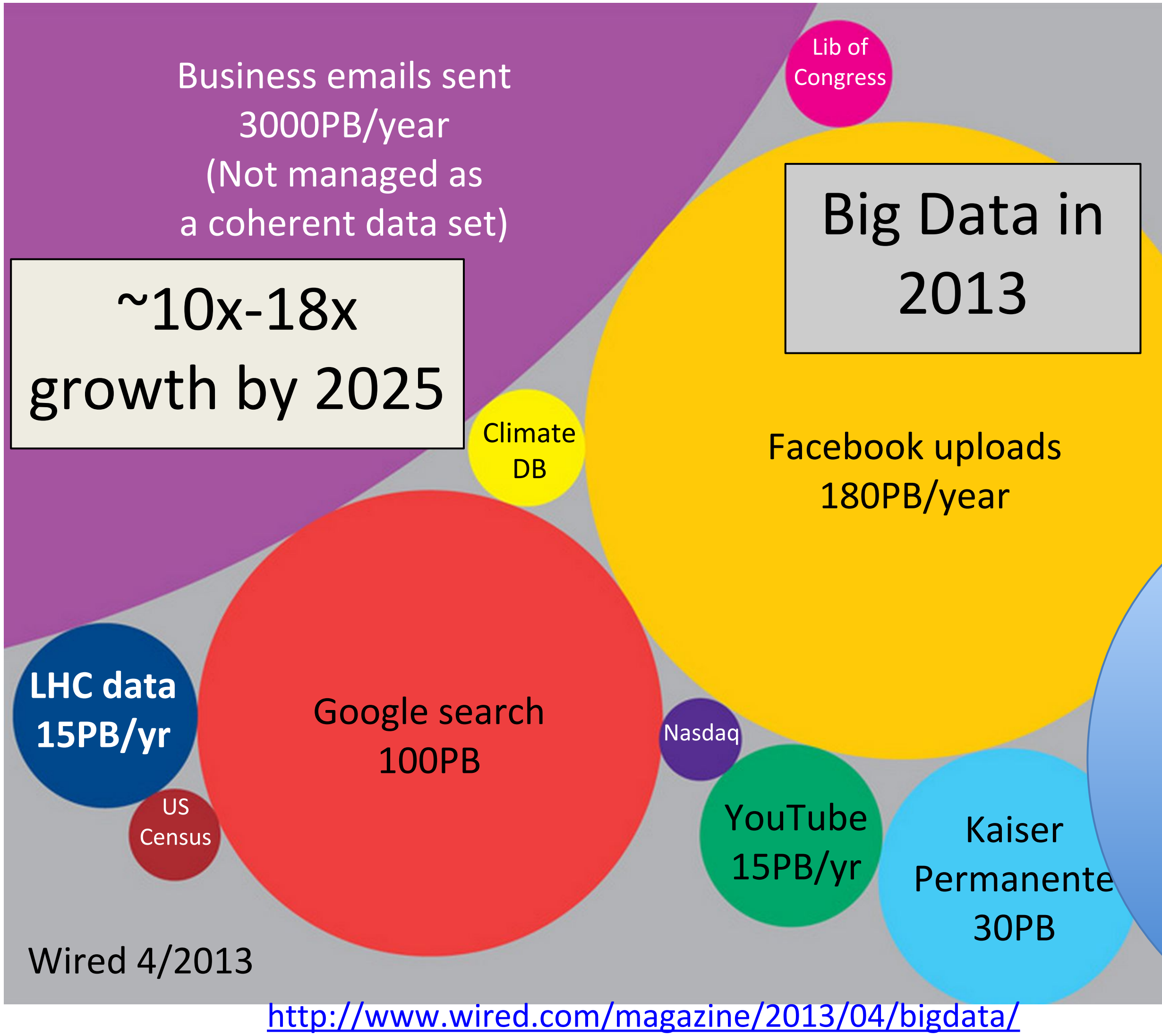
1 in 10,000,000,000,000

Like looking for a single drop of water from the Geneve Jet d'Eau over 2+ days

Jet d'Eau: 500 L/sec

ATLAS: ~1PB raw data/s off the detector filtered to 1-2 GB/s recorded

Big Data: Not a buzz word when it comes to ATLAS

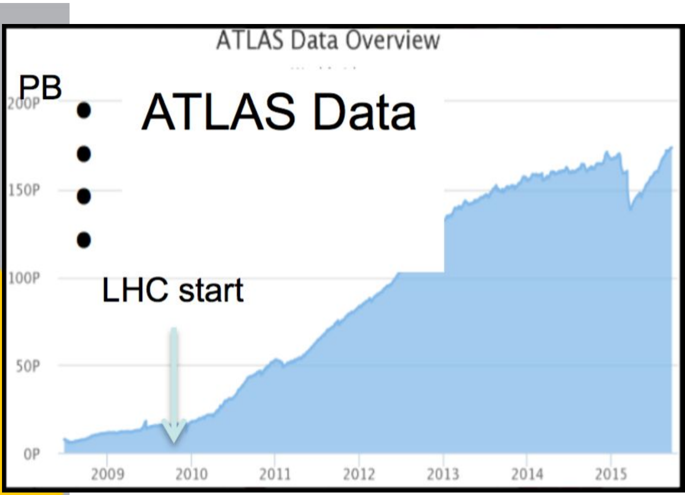


~10x-18x growth by 2025

Big Data in 2013

Wired 4/2013

http://www.wired.com/magazine/2013/04/bigdata/



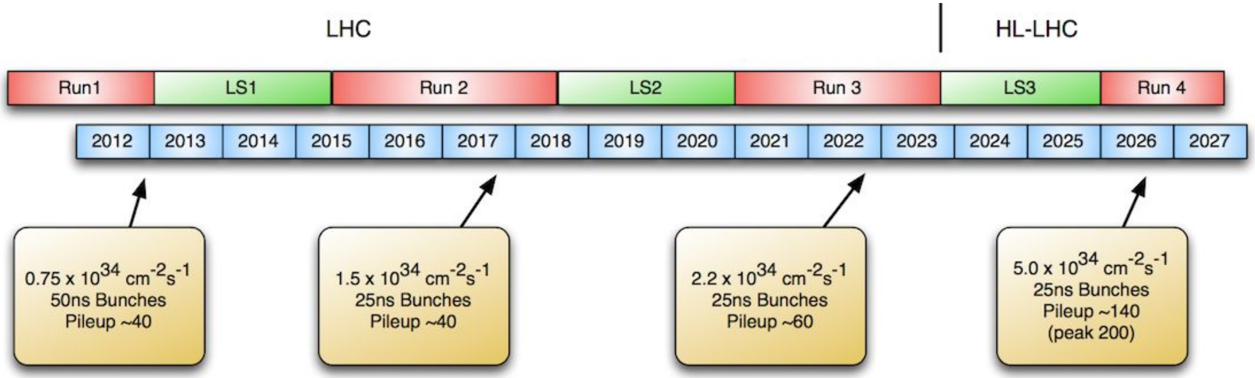
- ATLAS Data
-
-
-

Current ATLAS data set, all data products: 160+ PB

1M+ files moved/day

ATLAS Computing to 2025: the Outlook

	Integrated Lumi (fb
Run 1	25
Run 2	100
Run 3	300
HL-LHC	+300 per year



Storage: ~18x extrapolating today

~10x if we're smart

Fewer replicas, **use the network for remote cached data access**, **processing on demand**, more tape,...

Streaming is key to the smart path

	RAW (2 replicas)	Derived	Annual Total	Increase over now
Now	8PB/yr	x8	72PB	x1
HL-LHC do nothing	150PB/yr	x8	1350PB	x18
HL-LHC smart	150PB/yr	x4	750PB	x10

CPU: ~30x extrapolating today

~8x if we're smart

Simulation improvements, re-engineering software for concurrency, algorithmic improvements

Step	Approx. Fraction Today	HL-LHC do nothing multiplication factor	HL-LHC do nothing CPU increase	HL-LHC smart multiplication factor	HL-LHC smart CPU increase
Generation	0.05	20	1	5	0.25
Simulation	0.45	5	2.25	3	1.35
Digitisation	0.05	20	1	10	0.5
Reco (MC)	0.15	100	15	15	2.25
Reco (Data)	0.1	100	10	25	2.5
Analysis	0.2	10	2	5	1
Total (in units of today's compute)	**1**		**31.25**		**7.85**

Guesstimates from Graeme Stewart, CHEP 2015

ATLAS Computing Essentials

- Globally distributed, by necessity: computing follows the people and the support dollars
  - The ATLAS Grid would be about #27 on the HPC Top 500
  - And it isn't enough: big push into opportunistic resources
- 140+ heterogeneous sites sharing 160PB and processing exabytes per year, with a few FTEs of operations effort
- Our ability to do that is grounded in
  - **Excellent networking**, the bedrock enabler for the success of LHC computing since its inception
  - **Workflow management** that is intelligent, flexible, adaptive, and intimately tied to **dataflow management**
  - Dataflow management must minimize storage demands by **replicating minimally and intelligently**, using our **networks to the fullest** by sending **only the data we need, only where we need it**

From fine grained steering to fine grained data flow

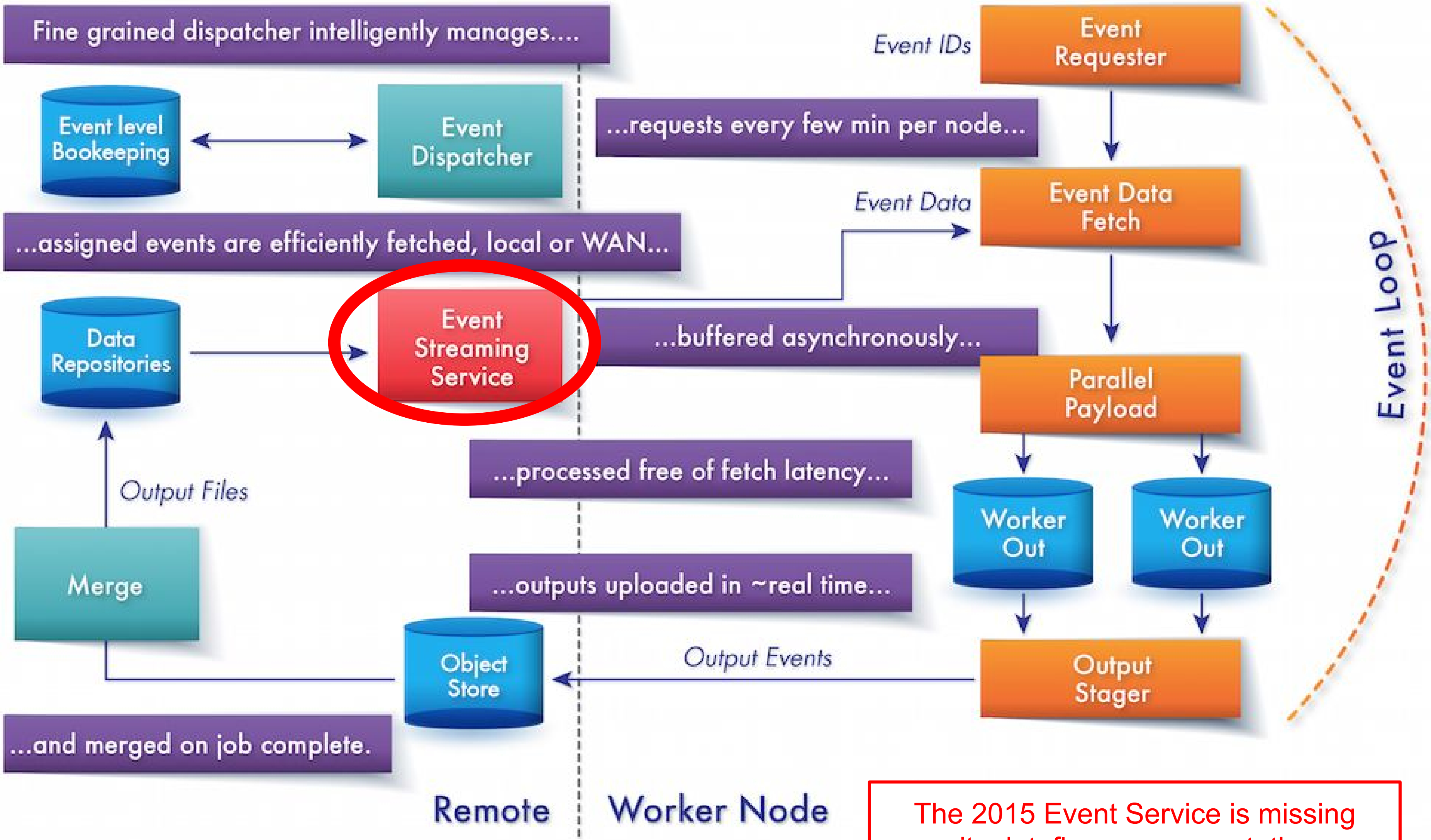
2015: ATLAS is commissioning the **Event Service,** a new approach to HEP processing: adaptive, fine-grained workflows for optimal use of opportunistic resources

- Agile, dynamic tailoring of workloads to fit the scheduling opportunities of the moment (HPC backfill)
- Loss-less termination (EC2 spot market node disappears)

2016+: The event service gives us fine-grained steering in the workflow, the next step is to do the same for the data flow, with streaming: the **Event Streaming Service**

- Efficient, intelligent distributed data access
- Maximizing the return on our excellent networks to minimize storage needs
- Paving the way to data on demand: virtual data

The Event Service 2015


Fine grained dispatcher intelligently manages....
Event level Bookeeping
Event Dispatcher
Event IDs
Event Requester
...requests every few min per node...
Event Data
Event Data Fetch
...assigned events are efficiently fetched, local or WAN...
Data Repositories
Event Streaming Service
...buffered asynchronously...
Parallel Payload
Event Loop
...processed free of fetch latency...
Output Files
Worker Out
Worker Out
Merge
...outputs uploaded in ~real time...
Object Store
Output Events
Output Stager
...and merged on job complete.
Remote
Worker Node

The 2015 Event Service is missing its dataflow component, the Event Streaming Service

The Event Streaming Service (ESS)

- The Event Service can integrate perfectly with a similarly event-level data delivery service, the ESS, that responds to requests for ‘science data objects’ by intelligently marshaling and sending the data needed
- The service can encompass
  - CDN-like optimization of data sourcing ‘close’ to the client
  - Knowledge of the data itself sufficient to intelligently skim/slim during marshaling
  - Servicing the request via processing on demand rather than serving pre-existing data (replacing storage with cheaper CPU cycles)
- We have to build it as an exascale system: we process today >1 Exabyte /year
- Currently at the design/prototyping stage

Building the ESS

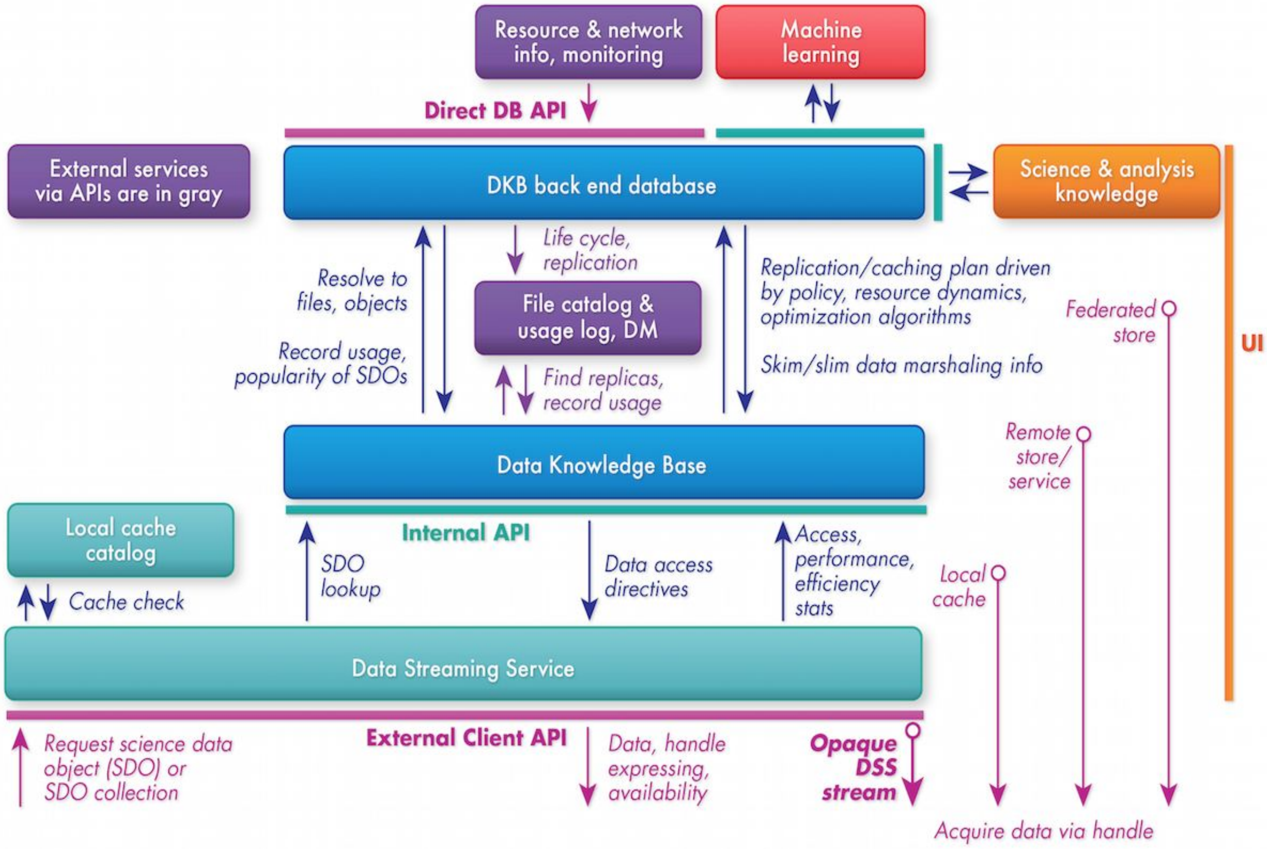
Two primary components:

Data Streaming Service

- CDN-like intelligence in finding the most efficient path to the data
- With minimal replication
- Data marshaling
- Smart local caching

Informed by the **Data Knowledge Base** providing the intelligence on

- Dynamic resource landscape
- Science data object (SDO) knowledge
- Analysis processes & priorities


Resource & network info, monitoring
Machine learning
Direct DB API
External services via APIs are in gray
DKB back end database
Science & analysis knowledge
Resolve to files, objects
Life cycle, replication
File catalog & usage log, DM
Record usage, popularity of SDOs
Find replicas, record usage
Replication/caching plan driven by policy, resource dynamics, optimization algorithms
Skim/slim data marshaling info
Federated store
UI
Data Knowledge Base
Remote store/ service
Local cache catalog
Cache check
Internal API
SDO lookup
Data access directives
Access, performance, efficiency stats
Local cache
Data Streaming Service
External Client API
Request science data object (SDO) or SDO collection
Data, handle expressing, availability
Opaque DSS stream
Acquire data via handle

On the lookout for tools with which to build these!

Finally

- ATLAS today pushes the bounds of data intensive science with exascale processing workflows on a 160 PB data sample across >100 global sites
- ATLAS is moving to a completely new processing model to sustain its science as its computing needs grow tenfold plus
  - Agile, fine grained processing harvesting resources opportunistically and delivering just-what's-needed data efficiently on a global stage
- The Event Service is Phase 1, deployed and delivering
  - 50k concurrent ATLAS simulation production slots on AWS EC2 spot market, 50k also on NERSC Edison, now spreading across the grid
- Now starting Phase 2: an Event Streaming Service that streams our Exabyte-scale data flows through the ES
  - Dynamic science-aware dataflows using the network heavily but efficiently, with smart caching and asynchronous processing
- Eager for tools to build it from CS, open source, commercial
  - These days we can build amazing things precisely attuned to our needs by plucking powerful tools of these shelves
  - We look forward to a rich toolset for exascale data streaming