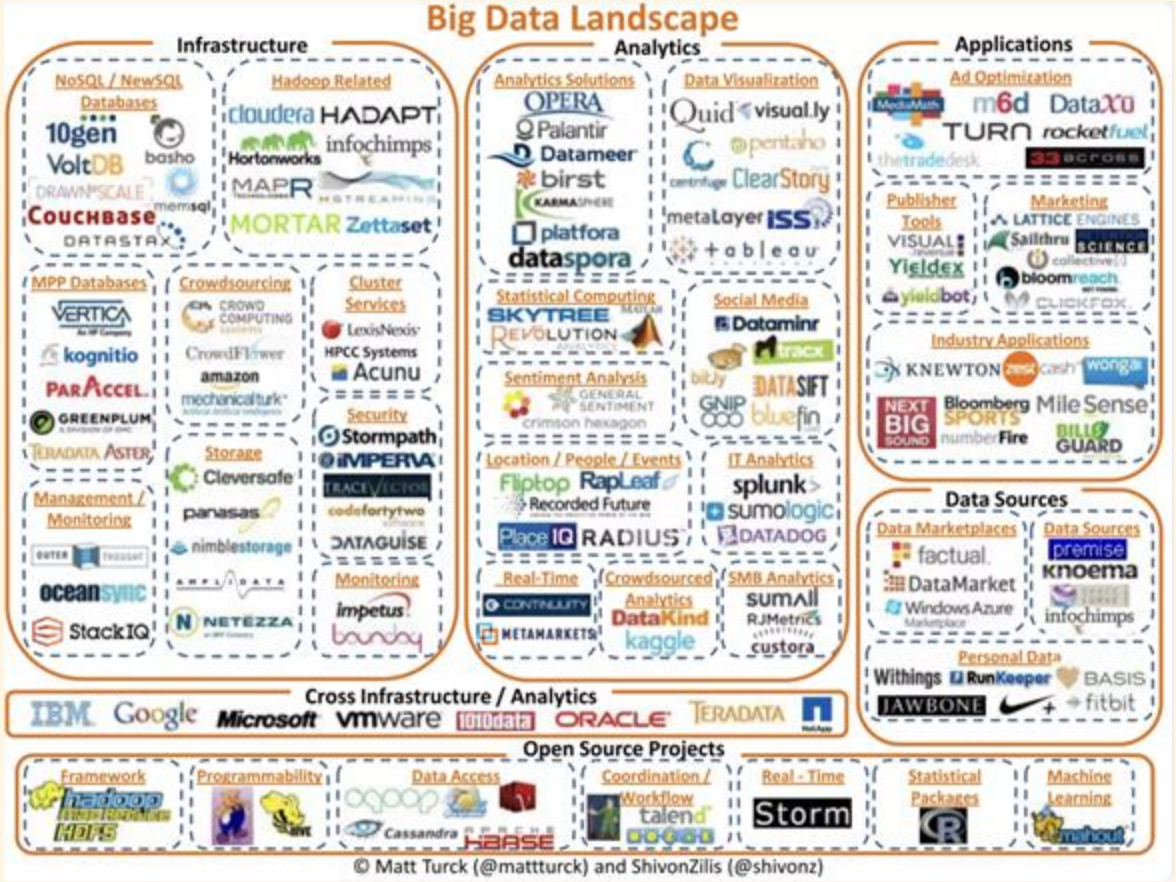
Everpure

Alpha Without the Overhead:

Simplifying the Modern Quant Data Pipeline

Rob Glanzman
Global Strategic Alliances Principal Architect, Financial Services
Everpure

Where we began, circa 2012



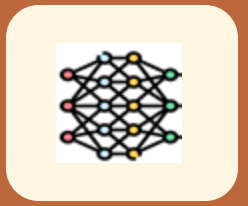
© Matt Turck (@mattturck) and ShivonZilis (@shivonz)

Importance of AI infrastructure

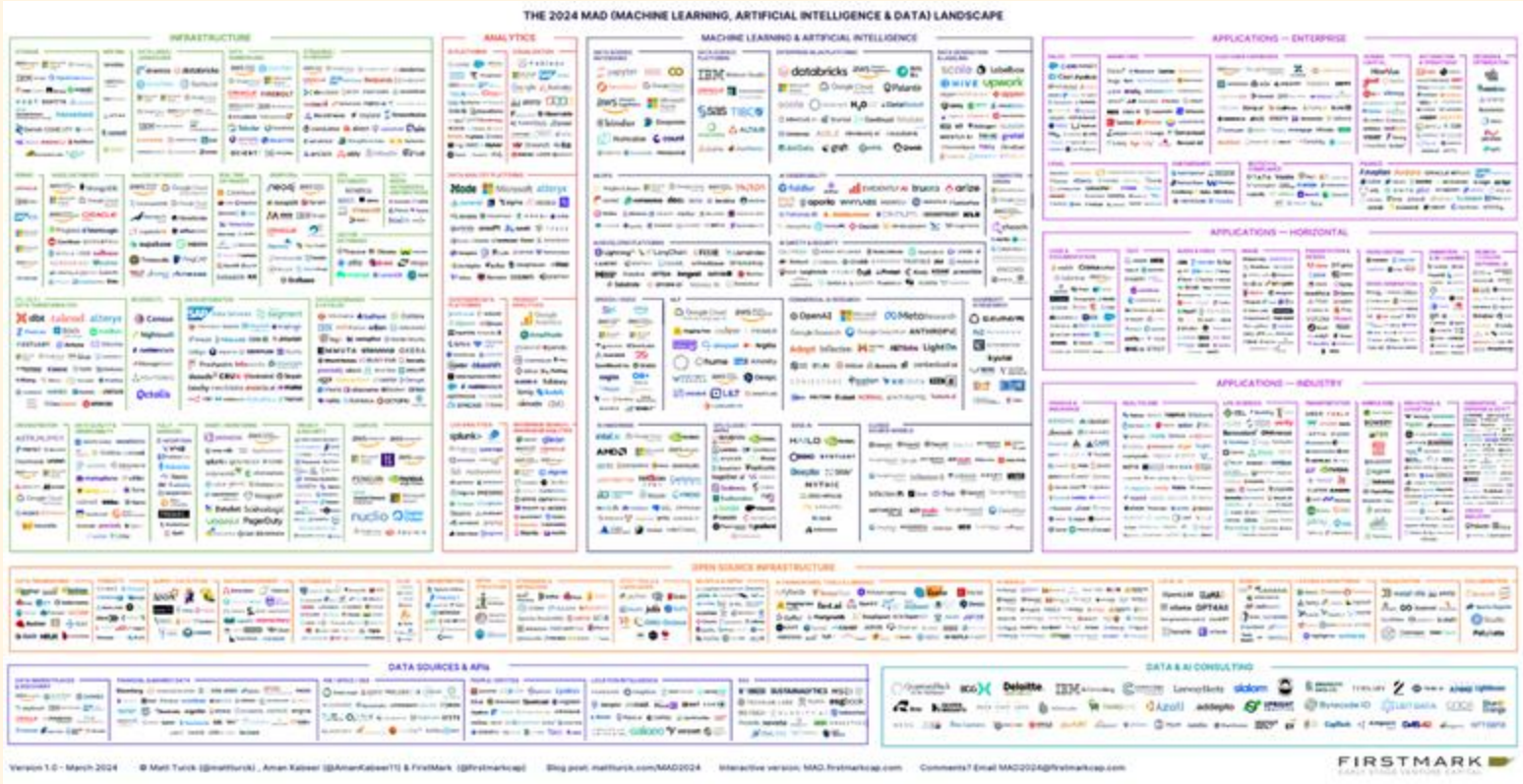
Considerations for data

Data curation

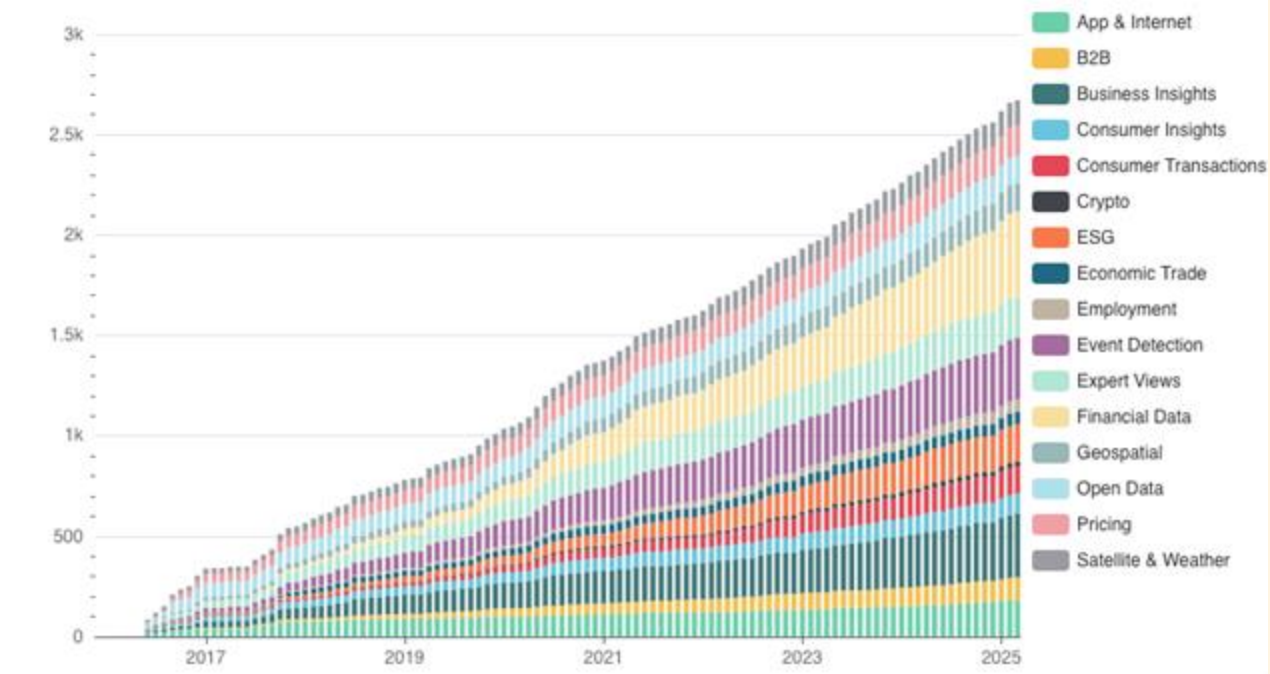
AI training and inference

Ingestion	Persistence	Processing	Analysis	Training	Inference
					
Sequential	Sequential	Sequential and random	Random access	Random access	Random access
Writes	Writes	Reads and writes	Reads	Reads and writes	Reads
Small and large files	Small and large files	Small and large files	Small and large files	Small files	Small and large Files
Varying concurrency	Varying concurrency	High concurrency	Low concurrency	High concurrency	High concurrency

AI Today…check back tomorrow



Examples of Data Used by Portfolio Managers



Source: /www.eaglealpha.com/what-is-alternative-data/

Importance of AI infrastructure

Considerations for data

Data curation

AI training and inference

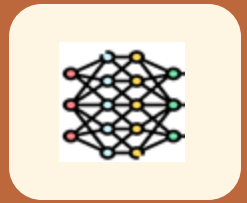
Ingestion

Persistence

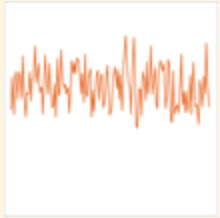
Processing

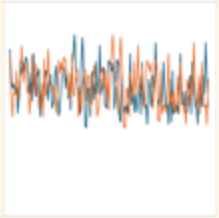
Analysis

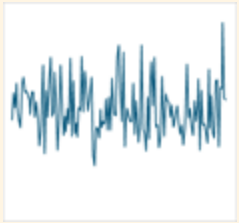
Training

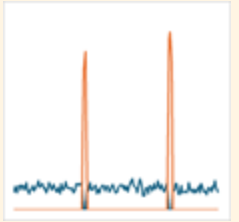


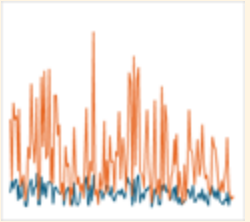
Inference











Importance of AI infrastructure

Considerations for data

Data curation

AI training and inference

Do these problems look familiar?

- Silos, data sprawl, security!
- Complexities
- Accessibility challenges
- Processing takes too long
- Storage and cloud costs
- ...

Importance of AI infrastructure

Considerations for data

Data curation

AI training and inference

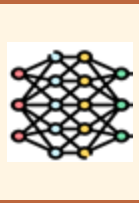
Ingestion

Persistence

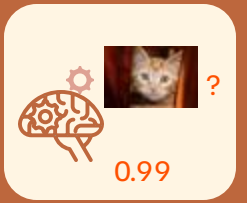
Processing

Analysis

Training



Inference



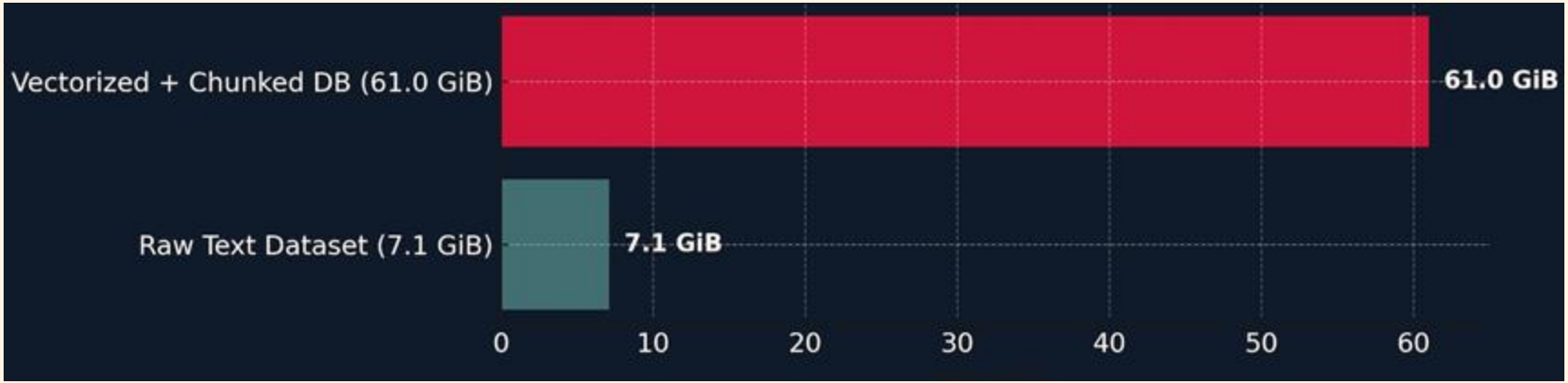
As you scale/
move into production

- Single data platform
- Radical simplification
- No tuning
- Predictable, performant expansion
- AI utility consumption model

Raw Embeddings Grow Fast... Really Fast

19101 text files from the Gutenberg project for a total raw volume of **7.1GB** resulting in just over **14 million embeddings over S3.**

8x (or 15x if we consider all db log files)!



Announcing Pure Key Value Accelerator (Pure KVA)

The Industry's First KV Cache Backend Supporting Both S3 and NFS Storage

Benefits

- High-performance caching system
- Seamlessly integrates with vLLM
- Production-ready with multi-GPU support

Core Innovation

- Supports both object (S3) and file (NFS)
- Transforms KV caches into shareable resources
- Enables cache reuse across your AI fleet

Key Achievements

- Up to 20x faster inference
- 50-70% storage savings
- Multi-GB/s throughput

Powering the Data Storage Platform for AI

Introducing FlashBlade//EXA™

Ultimate performance scale
Up to 10TB/s of multi-dimensional performance

Near unlimited scale
Exabyte scale & namespace

Industry leading TCO
Best price vs performance in the market

Next gen architecture
Disaggregated metadata / data

Predictable performance
Improved GPUs utilization

10TB/s ?

All SEC filings since 1994 are ~30 TB.

At 10 TB/s, //EXA can load all corporate disclosures…in three seconds!

Historical equities and options tick data for the U.S. since 2000 are ~500 TB.

At 10 TB/s, //EXA can load two decades of market microstructure…in 50 seconds!

Visa and Mastercard service ~150B transactions/year consuming ~2 PB of data.

At 10 TB/s, //EXA can reload an entire year of global payments…in 200 seconds!

Learn more!

Visit **Booth B7** to connect with the Everpure team.

Fireside Chat: Compute is the new Alpha: Building quant platforms that scale productivity, performance, and ROI

April 1st, 12:40 PM - 1:00 PM, TechX Stage

- [Moderator] Alan Russell, Former Lead Quant Engineer - Head of Front Office Engineering, Brevan Howard
- Rob Glanzman, Global Strategic Alliances Principal Architect, Financial Services, Everpure
- Ritesh Bansal, Managing Director, Risk Quant Platforms, Citi

White Paper: Quantitative Trading with Pure Storage Solutions