Real-Time Data Analysis and ML for Fraud Prevention

Lessons Learned from the Large Scale Decisioning Platform

Mikhail Kourjanski, Principal Architect

QCON London, March 2018

Introduction

What does FinTech do?

Did you pay for QCon registration with PayPal?

- Is it a Pmt Processor (buyer-seller)?
- Is it banking?
- Is it lending?
- Mobile platform?
- Merchant/Mktplace Integrator?
- … P2P Payments?
- … Global / Cross-border?
- … Multi-Brand?
- … Social Media?
- … Blockchain?

The answer depends on Product, Brand, Jurisdiction/Geo…

PayPal venmo xoom Braintree Paydiant

User Experience (UX) Defines FinTech Success

And what defines UX?

- Block fraud…
- …with low False Positives *(don't block good folks!)*
- Buyer and Seller Protections
- Customer data not shared with merchants
- Regulatory Compliance => Customer Safety

➢ PayPal brand promise starts with trust…

- Customer intent
- Sentiment
- Reduce friction:
  - Customer support and conflict resolutions
  - Onboarding
  - Offers that make sense

➢ And enhances UX from acceptance to delight

Consumers Trust PayPal to Help Protect Their Information and Transactions

Sources: Nielsen, Dept of Commerce, JP Morgan; PayPal & IPSOS Study Feb 2015; Symantec, Gemalto, LexisNexis

UX Delivered with Velocity and Scale

Supported by multi-billion dollar decisions everyday

~227MM
active users

~$1.25 Billion
TPV/day

~250MM
Fast Decisions/Day

~25 Billion
Computations/day

~60 Billion
Queries/day

~150 PB Data

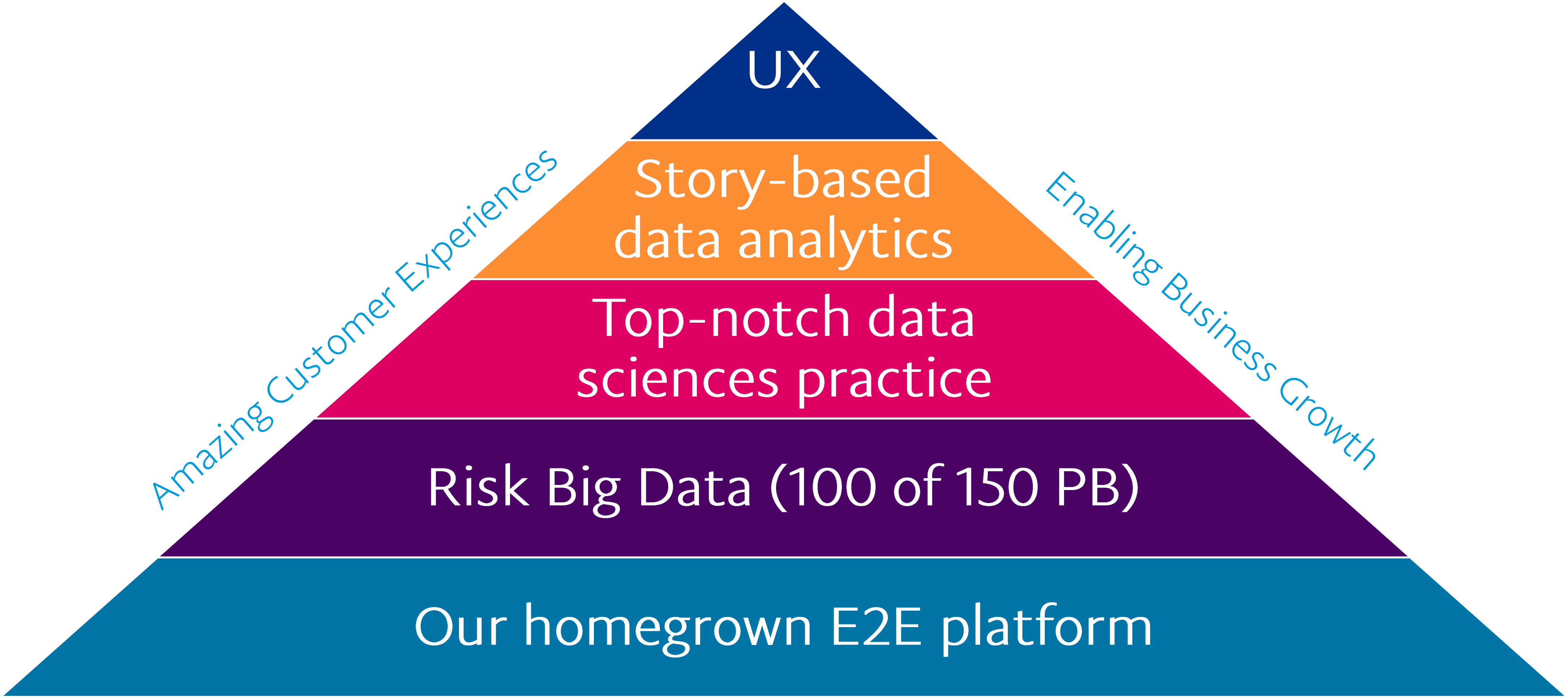
Data Scientist's Dream

Facts and numbers:

- PayPal - in more than 200 countries and regions.
- Secure Payments: $451 Billion global transaction volume in 2017
- Incoming fraud pressure markedly exceeds company revenues
- Sophistication of the modern day hacker attacks: distributed; high-velocity
- Compliance and Privacy: AML, Prevention of prohibited activities, KYC, PII protection

Risk Decisioning is a Competitive Advantage for PayPal

Key Differentiators Capabilities



Business Use Cases:
Where Exactly Decisioning is Used?

Carrying Risk of Transactions: Decisions at Checkpoints

Each payment transaction is a customer's story

Explore → Enroll – New Acct → Manage via Self-Service → Transact → Resolve

Reporting, and Analytics - Research

Explore
- Behaviour
- Offers
- Guest checkout

Enroll – New Acct
- Do we know you?
- Validations
- Credit Risk

Manage via Self-Service
- Login / Auth
- Wallet
- Profile

Transact
- Purchase <Buyer, Seller>
- Deposit
- Send Money
- Withdrawal

Resolve
- Complaints
- Chargebacks
- Recover NSF
- Investigations

Velocity

Linked Objects & Activities

Merchants: a World of Additional Checks

Representative Merchant evaluation criteria

- Merchant –Business Solvency
- Customer satisfaction - *Items Not Received; Significantly Not As Described*
- True Industry; Prohibited Goods - *Merchant Category Codes - MCC*
- Revenue rate of change: $\frac{d\,\mathbf{Revenue}}{\mathbf{dt}}$; *fast growth / wild fluctuations?*
- Linking; Compliance – AML (collusion)
- *Partnership / Marketplace - specific*

Types of Data at Checkpoints

What Data Do We Process?

Types of data affect choice of modeling methods and frameworks

Explore → Enroll – New Acct → Manage via Self-Service → Transact → Resolve

Structured data…

- *Numbers*
- *Strings*
- *Dates*
- *Geo, …*

Features

… + Unstructured data

- *Text – emails, customer interaction records*
- *ChatBot*
- *Voice - IVR*
- *Images*
- *Social media*

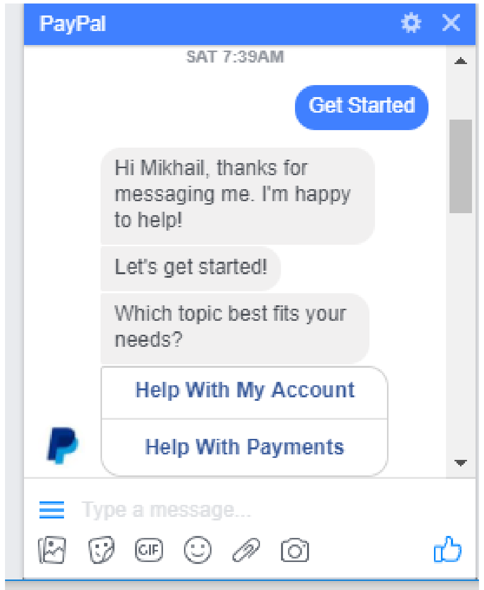
Machine Learning Use Cases

Structured data

- Checkpoint decisions (e.g. Transaction)
- Real-Time data is (usually) structured

Natural Language Processing - NLP

- Text analysis: customer interaction log; emails; extended Tx data
- IVR – voice-to-text
- Chat Bot



Challenge: multi languages

Image Recognition

- Merchant True Industry
- Marketplace New Listings – prevent Prohibited Goods

Opening the Hood:
Key Ingredients of Decision Making Process

A Story of a Payment: Serving Decisions at Checkpoints

Decisioning flow

Y/N, or Action
Decision for a Checkpoint

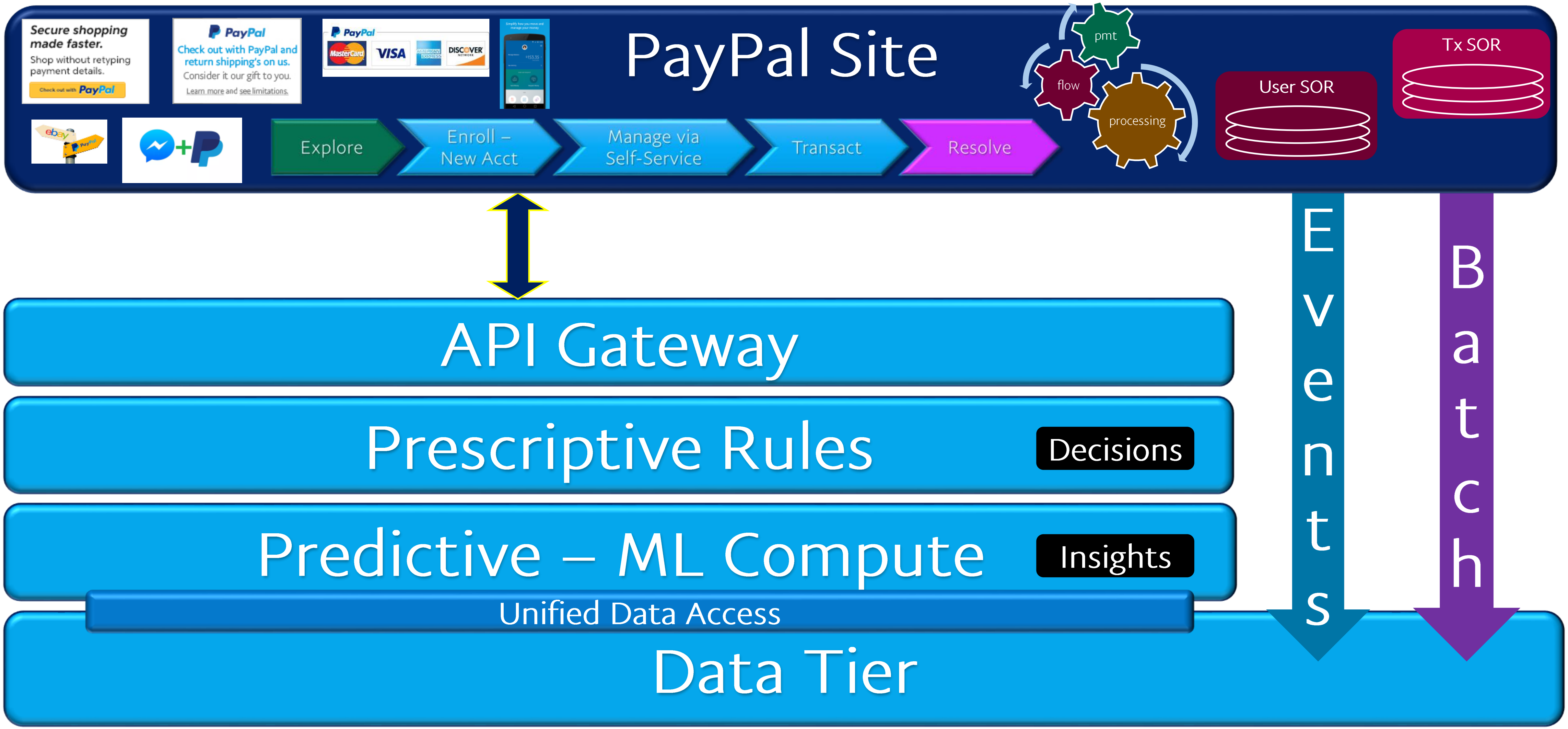
~75% calls at < 50ms;
deep inspections can be ~500ms

Decisioning Platform (Magic)

Fail-Open or Fail-Close? – ask Biz & Compliance

The Anatomy of Decisioning

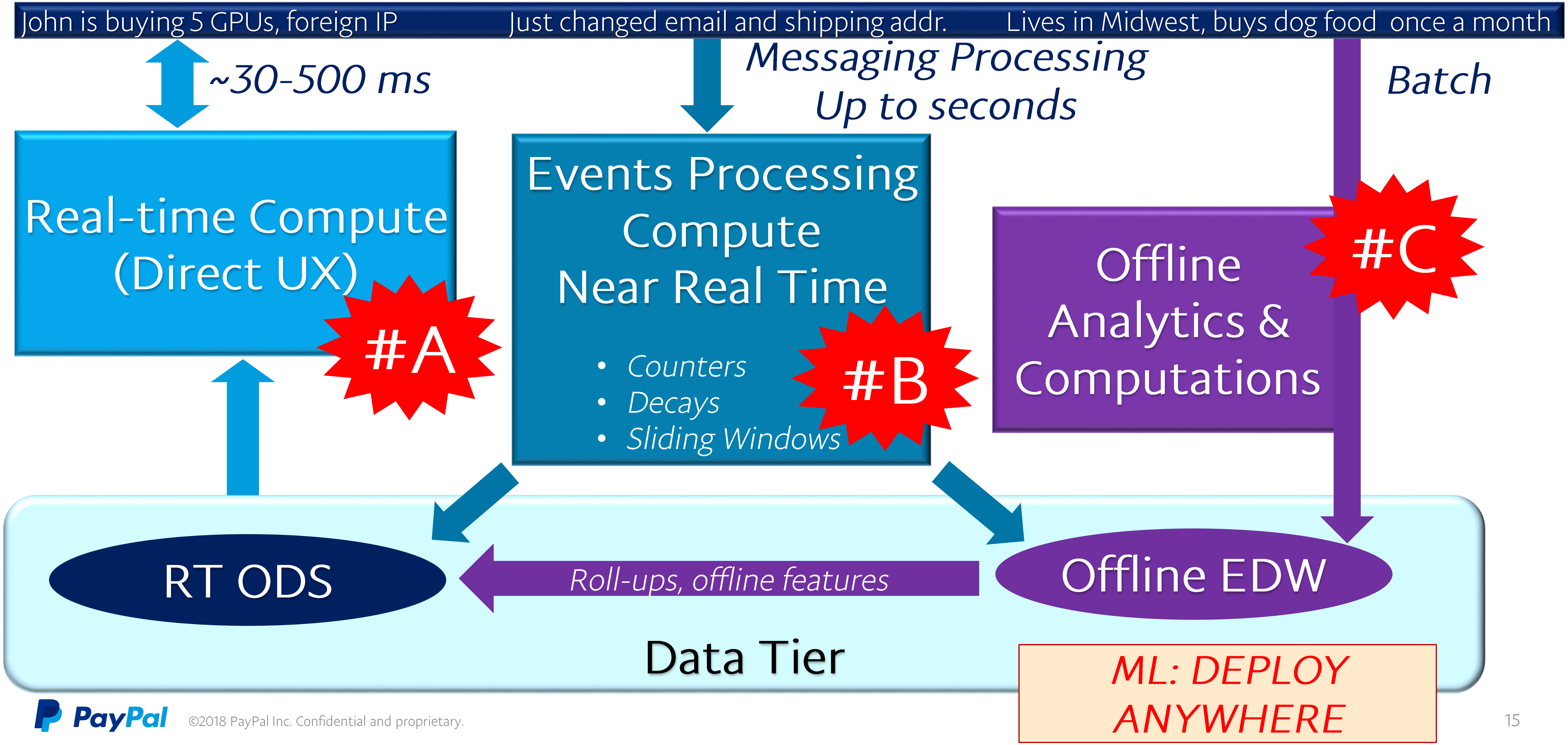
Decisioning flow



PayPal

Three Velocities of the Data Flows

Where to execute ML models (Inference) – in #A, or #B, or #C? ...And where to train?

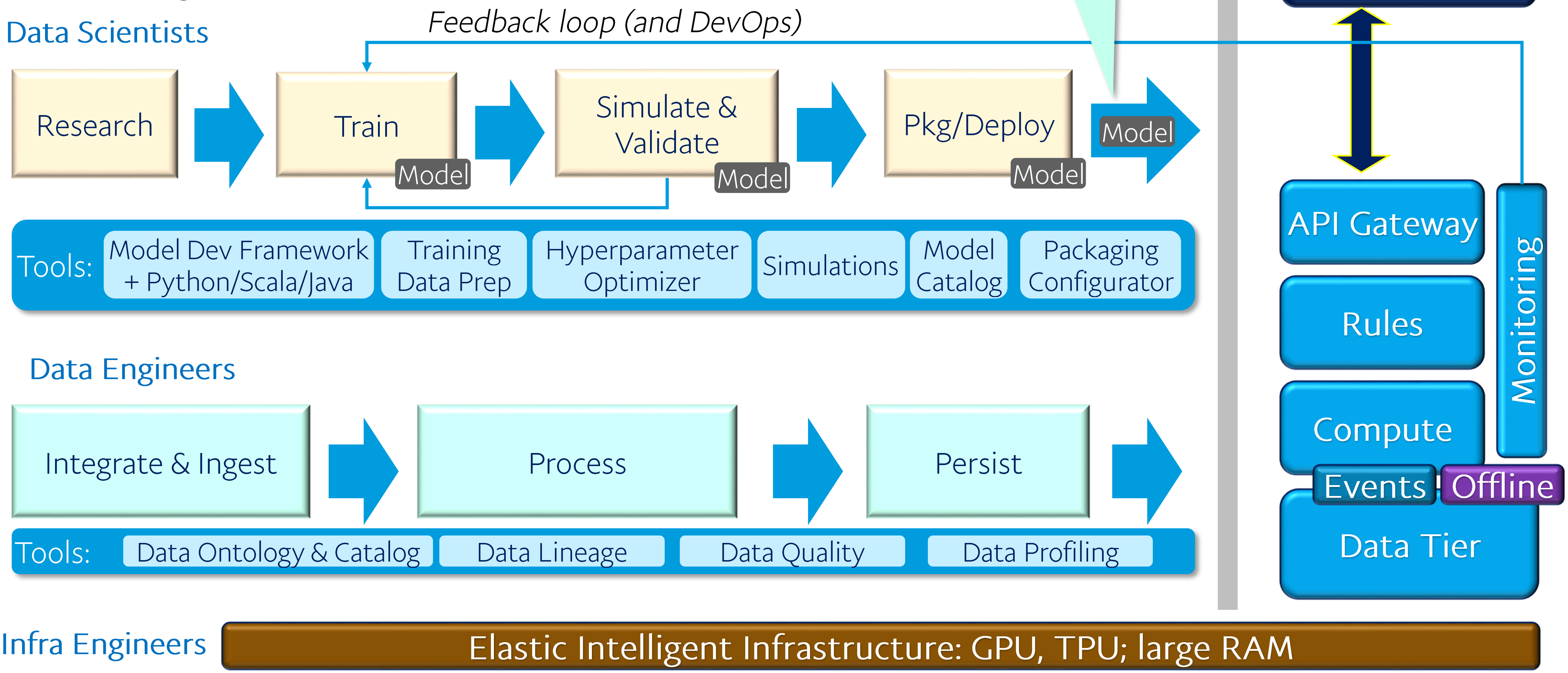


Ok, we got a Production stack.

But… Where is the Machine Learning?

ML Pipeline

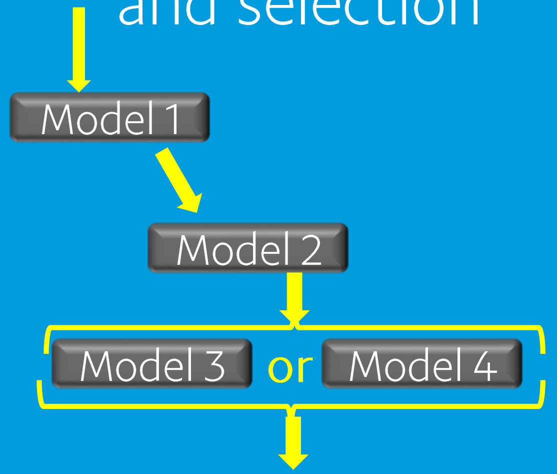
Cross-Organizational view



See Google at NIPS'15: "Hidden Technical Debt in Machine Learning Systems"

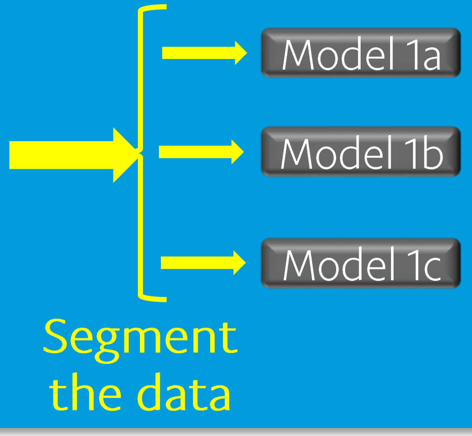
ML Models – Inferencing in Production Ecosystem

Model Composition
- Model sequencing and selection

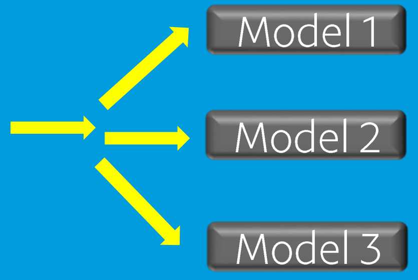


Avoid "Corrective Cascading"

Segment Model
- Models on different subsets of data



Model Ensemble
- Different models on same data



- ➢ Continuous comparative analysis of models' performance
- ➢ => faster feedback loop

- ➢ Deploy any model from portfolio, incl. multiple versions of same model

Variety of Modeling Methods

Different methods require different lifecycle and compute model

Real-Time Interactivity for UX

Regression; Trees; Rules

- *Fresh fraud – limited vol. sample data*
- *Simple compute formula (rules)*

Artificial Neural Nets

- *E.g. Acct Takeover models*
- *Large training data set*
- *Model segments, ensembles*

Canonical development pipeline

NRT Graph Compute

- *Paypal Legacy: ~2bn nodes, NRT*
- *In Dev: iterative compute; much larger size*

Deep Learning

- NLP, images
- Online learning

Unsupervised

- Anomaly detection

Internal state; and may have Compute, Training in Live / NRT

How to Manage Data?

Data Tier

Types of data stores

Data Stores

Real-time ODS	Near-Real Time Streaming; Big Data NoSQL	Enterprise Data Warehouse
ORACLE, AEROSPIKE	Spark, Apache Kafka, Apache HBase	hadoop, TERADATA

- NoSQL in-mem: < 1ms at 95%; < 4ms at 99^{th}. Not real ACID, not SOR => rigor to restore readiness and redundancy
- Oracle: 24 nodes main cluster - ACID, RDBMS
- ~30 billion queries/day (Decisioning)

> How?? – Big Data at (relatively) fast speed
  - *Key space*
  - *Read or Write optimized?*

> What technology to choose??

- 5 billion messages/day through Kafka (online to offline data flow), which is 100 to 150 TB of data
- ~100pb data in DW

Cloud Appeal, but Beware of Compliance, Privacy.

Data Tier

Fast Data and Big Data

Data Stores

Real-time ODS	Near-Real Time Streaming; Big Data NoSQL	Enterprise Data Warehouse

1% data vol
(~1pb)

Blurring the Line Between Fast and Big:

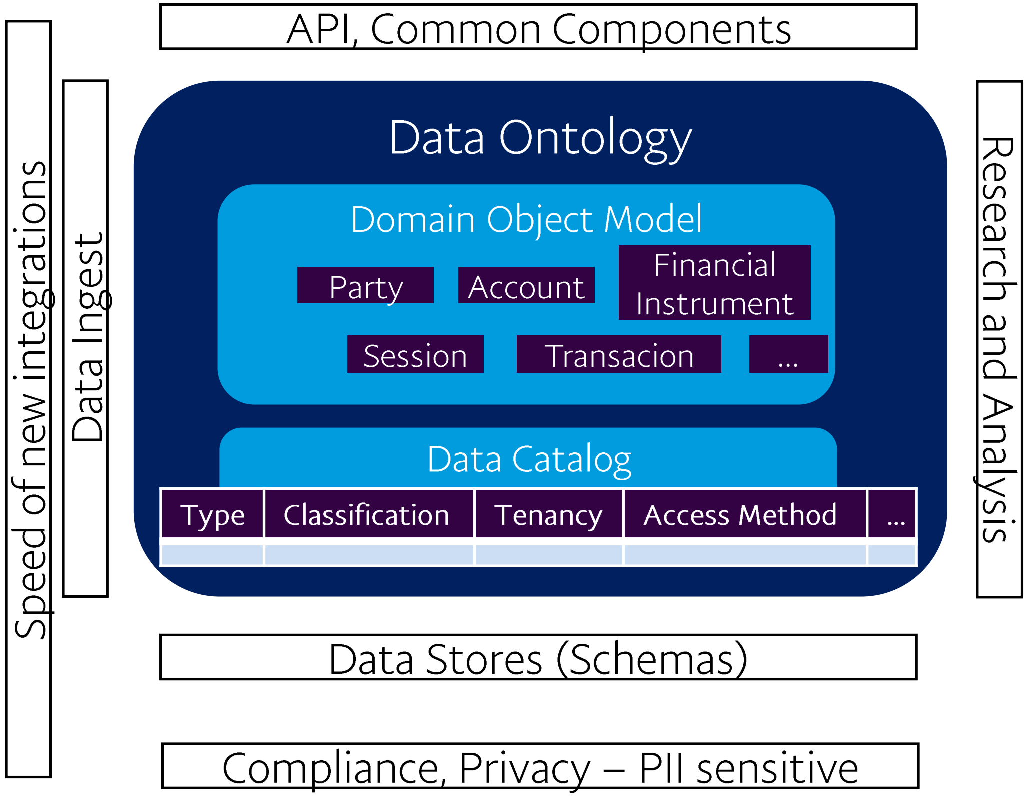
- Explicit class of NRT Data Stores
- Need vast raw data for DL
- Therefore, shifting historically offline systems (NoSQL) into NRT

99% data vol
(~100pb – Risk data)

Making more raw data available to NRT Compute

Data Management Discipline

FinTech Rigor for Compliance, Security and Privacy



- Know your data: Ontology and Catalog. Classification.
- Make it part of the Engineering/Delivery process
- Tokenize PII
- Where is data stored? – geo/jurisdiction, and Lineage
- Data Quality: technical; business

Parting Thoughts

Takeaways

- Architect for success
- Agnostic to Framework/Language/Product
- Automate and hide complexity - *aaS
- Know your data and all its flows
- Provision for Infrastructure elasticity (consider cloud)
- Smart Platform – build intelligent skills in platform
- $ can add up quickly - Pay close attention to TCO

The journey continues ...

Thank You!

Mikhail Kourjanski
Principal Architect

Email: mkourjanski@paypal.com

Linkedin: https://www.linkedin.com/in/mikhail-kourjanski-79358/