

AWS

S U M M I T

Building Your First Data Lake

Modern Data Architectures on AWS

Dickson Yue
Solution Architect
21 June 2017



Today's conversation



Business drivers for a Data Lake



Designing and building



Production use cases

Business Outcomes on a Modern Data Architecture



Outcome 1 : Modernize and consolidate

- Insights to enhance business applications and create new digital services



Outcome 2 : Innovate for new revenues

- Personalization, demand forecasting, risk analysis



Outcome 3 : Real-time engagement

- Interactive customer experience, event-driven automation, fraud detection

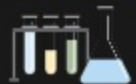


Outcome 4 : Automate for expansive reach

- Automation of business processes and physical infrastructure

Expanding access requirements

Data
scientists



Data
analysts



Business
users



Engagement
platforms



Automation /
events



1. More personas need access to data, through appropriate tools
2. More systems need to link to data for decision and process automation
3. Users need to be able to find information, and access it securely

Exponential growth of business data

Transactions



ERP



Web logs /
cookies



Connected
devices



Social media



1. Data must be captured from diverse sources at speed and scale
2. Data needs to be pulled together, breaking down traditional silos
3. Benefits need to far outweigh the costs of collection and analysis

Modern data architecture

Insights to enhance business applications, new digital services

Data
sources

Ingest

Scale (Batch)

Serving

Speed (Real-time)

Data scientists



Data analysts



Business users



Engagement platforms



Automation / events



Modern data architecture

Insights to enhance business applications, new digital services

Data sources

Transactions



ERP



Web logs / cookies



Connected devices



Social media



Ingest

Scale (Batch)

Speed (Real-time)

Serving

Data scientists



Data analysts



Business users



Engagement platforms

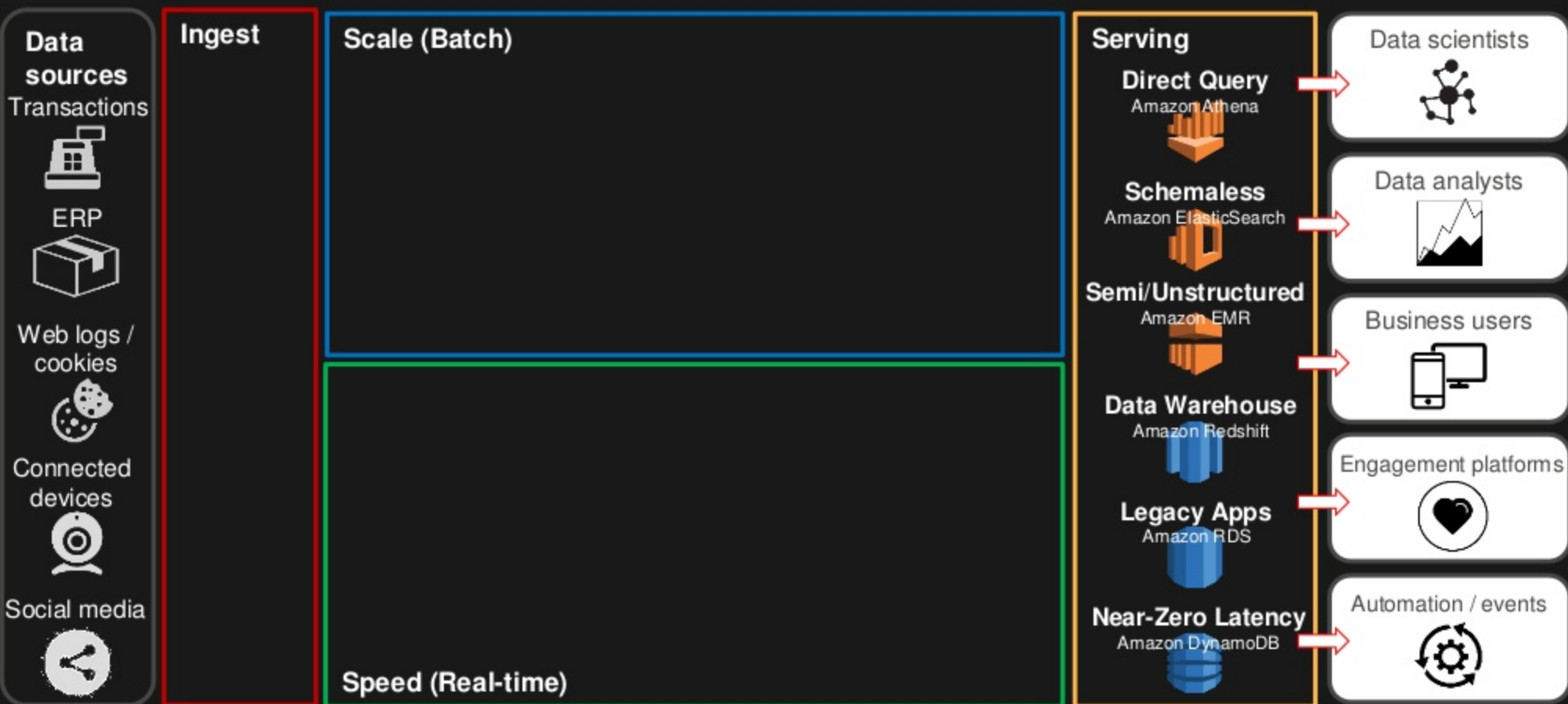


Automation / events



Modern data architecture

Insights to enhance business applications, new digital services



Characteristics of a Data Lake



Collect
Anything



Dive in
Anywhere



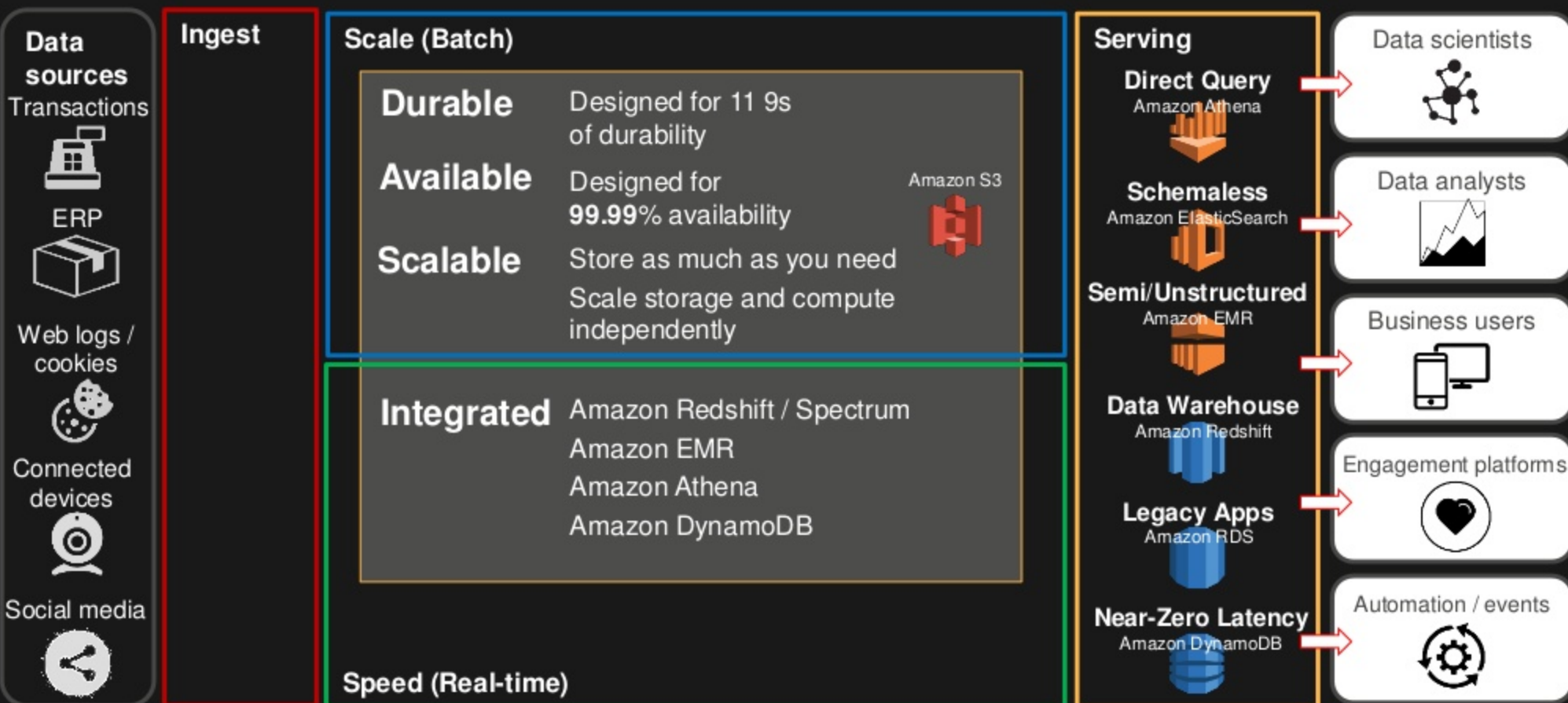
Flexible
Access



Future
Proof

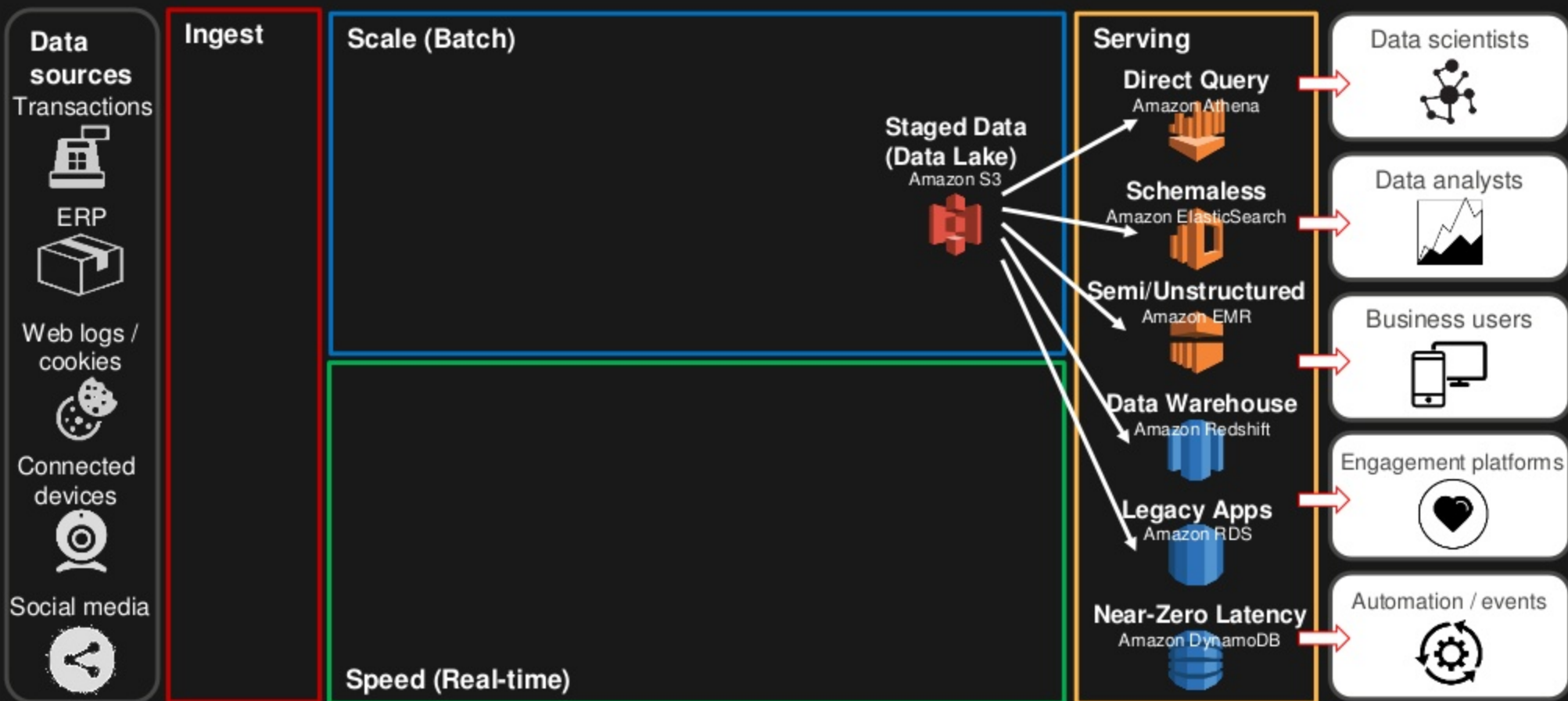
Modern data architecture

Insights to enhance business applications, new digital services



Modern data architecture

Insights to enhance business applications, new digital services



Today's conversation



Business drivers for a Data Lake



Designing and building



Production use cases

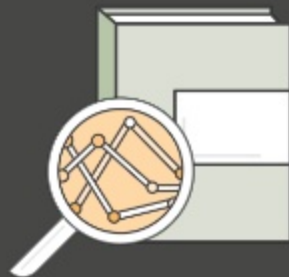
Important Components of a Data Lake



Access &
User Interface



Ingest & Store



Catalogue
& Search



Protect
& Secure

Data Ingestion into S3



AWS Direct Connect



S3 Transfer
Acceleration



Amazon Kinesis
Firehose



AWS Storage
Gateway



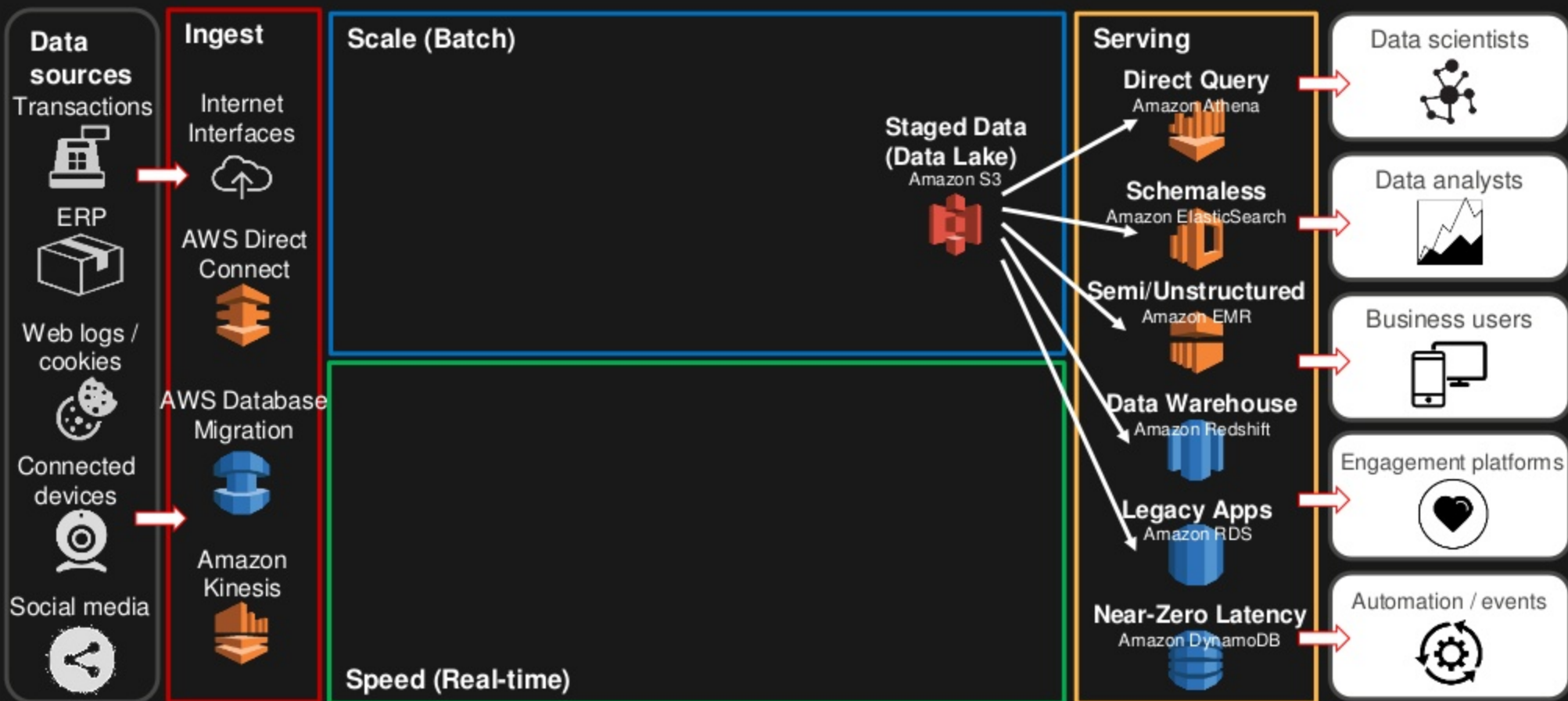
ISV Connectors



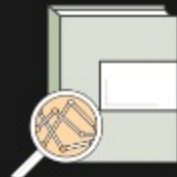
AWS Snowball

Modern data architecture

Insights to enhance business applications, new digital services

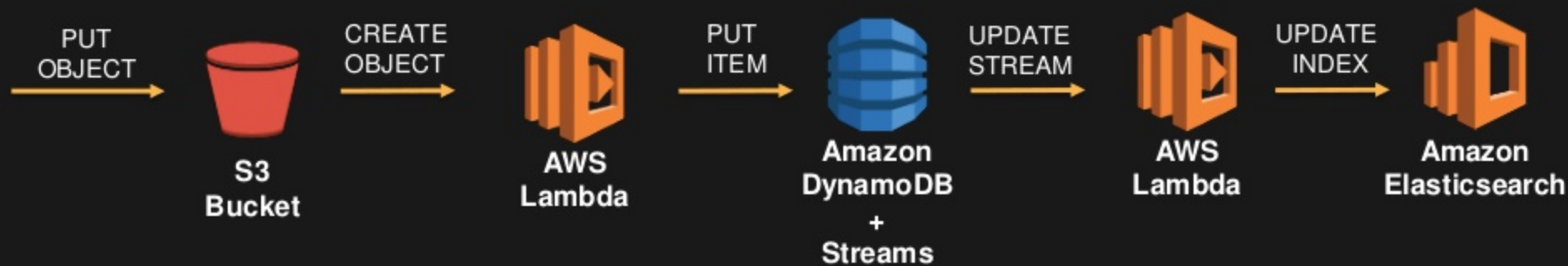


Building a Data Catalogue



- Aggregated information about your storage & streaming layer
- Storage service for metadata
 - Ownership, data lineage
- Data abstraction layer
 - Customer data = collection of prefixes
- Enabling data discovery
- API for use by entitlements service

Populating Metadata and Search



Available 2H 2017



**AWS
Glue**

Managed Transform Engine

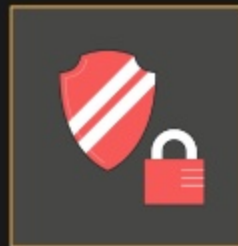
Job Scheduler

Data Catalog

Built on Apache Spark

Integrated with S3, RDS, Redshift & any
JDBC-compliant data store

Implement the right cloud security controls



Encryption

- SSL endpoints
- Server Side Encryption (SSE-S3)
- S3 Server Side Encryption with provided keys (SSE-C, SSE-KMS)
- Client-side Encryption



Security

- Identity and Access Management (IAM) policies
- Bucket policies
- Access Control Lists (ACLs)
- Private VPC endpoints to Amazon S3
- Pre-signed S3 URLs

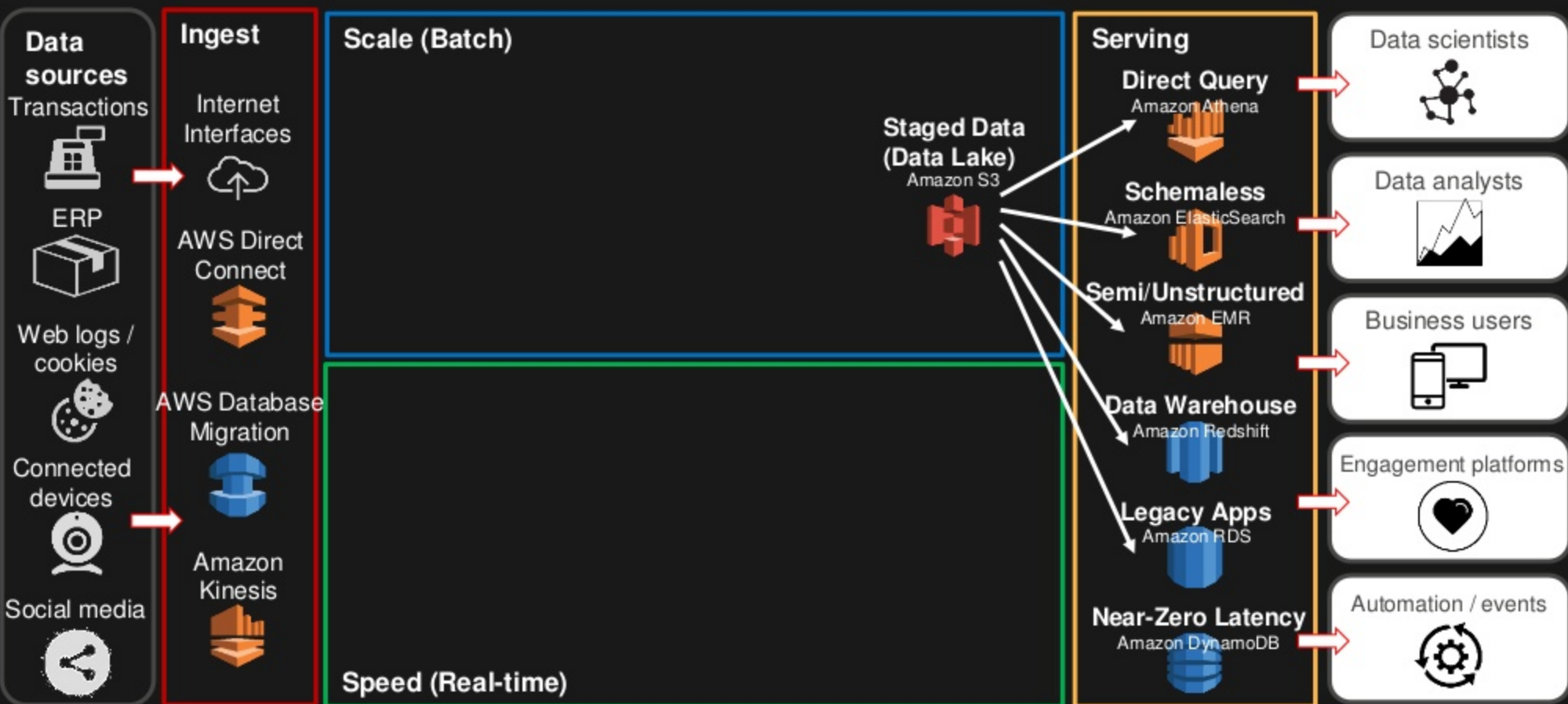


Audit & Compliance

- Buckets access logs
- Lifecycle Management Policies
- Versioning & MFA deletes
- Certifications – HIPAA, PCI, SOC 1/2/3 etc.

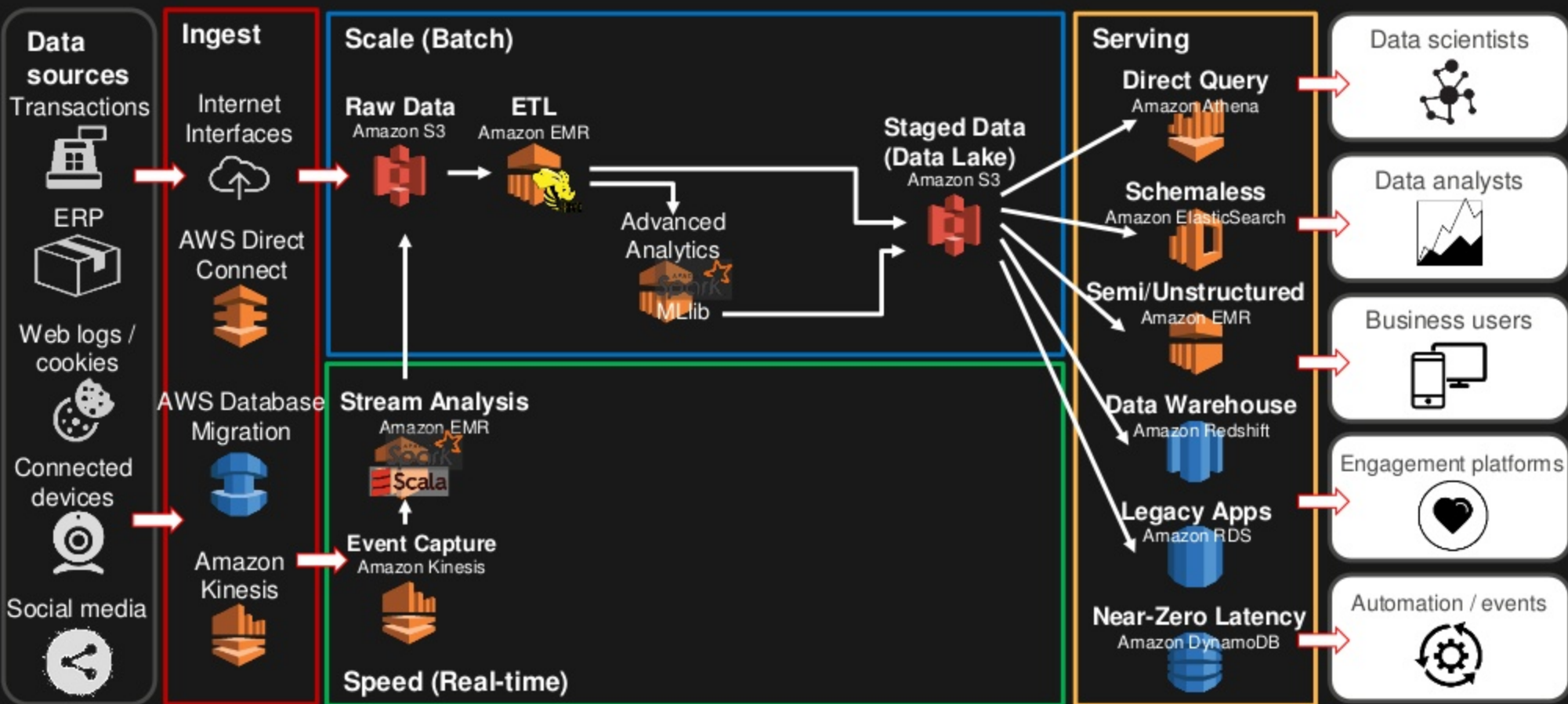
Modern data architecture

Insights to enhance business applications, new digital services



Modern data architecture

Insights to enhance business applications, new digital services



Today's conversation



Business drivers for a Data Lake



Designing and building



Production use cases

AWS

S U M M I T

Clickstream Analytics Pipeline

HK01

Angus Tse
Director of Engineer
21 June 2017



【圖輯】33度晴空下 汗水換來乾淨街道 清潔工玉姐的故事

她的汗水和努力，讓街坊可以享受清潔的街道。嚴夏之中，她辛勞的工作，你又知道嗎？



【圖輯】33度晴空下 汗水換來乾淨街道 清潔工玉姐的故事

【英國大選·圖輯】雙城記：兩個肯辛頓 看大選前夕分裂的...

【攝影集】攝影師游走各國遊樂場 拼出相似卻不同的童年時光

【紐約影像】紐約街頭紀實 公共空間光與影的偶遇

【社區影像】香港最後的人力車夫 時代轉變客人難求（有片）

娛樂



熱門文章

1 【廣東邕倫常案】全天候照顧病妻 殷妻翁曾向胞弟透露「好辛苦」



2 新西醫員審訊

Growth



DATA BEATS EMOTIONS

Sean Rad

Founder & CEO Tinder

Clickstream Analytics

Google Analytics (GA)

- Free and easy
- Excellent for initial
- Good learning materials

Google Analytics (GA)

- Free and easy
- Excellent for initial
- Good learning materials
- free version latency & accuracy issue
- GA 360 (Premium) + BigQuery are expensive
- Not flexible enough

Our needs

- Large data volume
- Raw data for Machine Learning
- Flexible for further processing
- Low latency

Building a scalable pipeline on AWS

Piwik

- Open-source analytics platform
- Realtime dashboard
- Web & mobile SDK
 - PageView
 - Content / Media
 - A/B Test

PIWIK

Finding the Piwik Tracking Code

To use all the features described in this page, you need to use the latest version of the tracking code. To find the tracking code for your website, follow the steps below:

- log in to Piwik with your admin or Super User account
- click on your username in the top right menu, and click *Settings* to access the administration area
- click on *Tracking Code* in the left menu
- copy and paste the JavaScript tracking code into your pages, just after the opening `<body>` tag (or within the `<head>` section)

The tracking code looks as follows:

```
<!-- Piwik -->
<script type="text/javascript">
var _paq = _paq || [];
_paq.push(['trackPageView']);
_paq.push(['enableLinkTracking']);
(function() {
var u="//(SOME_URL)";
_paq.push(['setTrackerUrl', u+'piwik.php']);
_paq.push(['setSiteId', (SITEID)]);
var d=document, g=d.createElement('script'), s=d.getElementsByTagName('script')[0];
g.type='text/javascript'; g.async=true; g.defer=true; g.src=u+'piwik.js'; s.parentNode.insertBefore
(g,s);
})();
</script>
<!-- End Piwik Code -->
```


Piwik

- Open-source analytics platform
- ~~Realtime dashboard~~
- Web & mobile SDK
 - PageView
 - Content / Media
 - A/B Test

PIWIK

Finding the Piwik Tracking Code

To use all the features described in this page, you need to use the latest version of the tracking code. To find the tracking code for your website, follow the steps below:

- log in to Piwik with your admin or Super User account
- click on your username in the top right menu, and click *Settings* to access the administration area
- click on *Tracking Code* in the left menu
- copy and paste the JavaScript tracking code into your pages, just after the opening `<body>` tag (or within the `<head>` section)

The tracking code looks as follows:

```
<!-- Piwik -->
<script type="text/javascript">
var _paq = _paq || [];
_paq.push(['trackPageView']);
_paq.push(['enableLinkTracking']);
(function() {
var u="//(SOME_URL)";
_paq.push(['setTrackerUrl', u+'piwik.php']);
_paq.push(['setSiteId', (SITEID)]);
var d=document, g=d.createElement('script'), s=d.getElementsByTagName('script')[0];
g.type='text/javascript'; g.async=true; g.defer=true; g.src=u+'piwik.js'; s.parentNode.insertBefore
(g,s);
})();
</script>
<!-- End Piwik Code -->
```

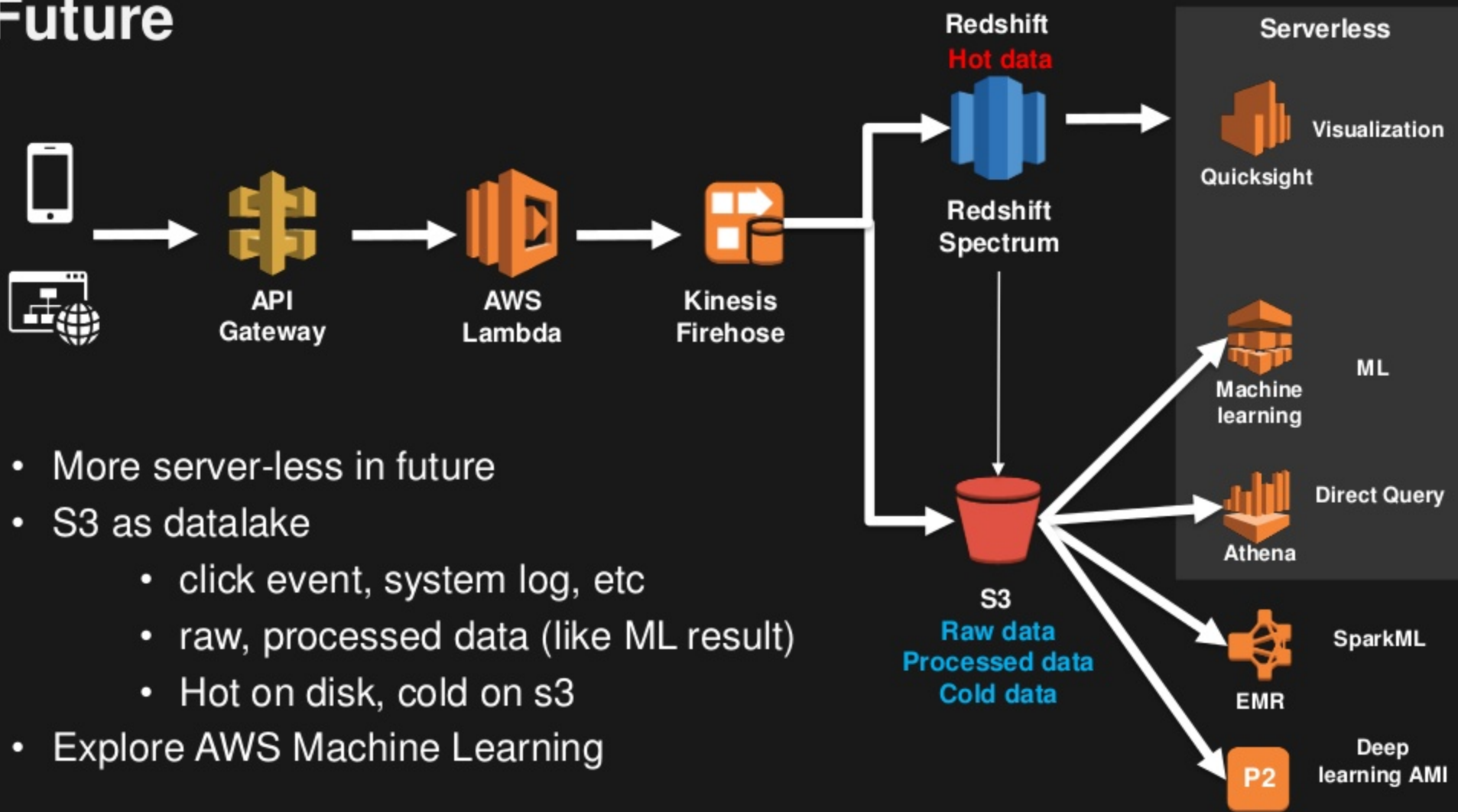
Phase 1



Experience on AWS

- **Complete** and Integrated
- **Quick**. 2 man weeks for first version
- **Easy** to scale
- **Minimal** maintenance cost

Future



Download HK01



We're Hiring



AWS

S U M M I T

Thanks

Join Us : hk01.com/job



Summary

1. S3 as data lake
2. Pick the right tool to match the persona requirements
3. Go serverless

