zyte

Web Data Maturity Model

September | 2022

Speaker

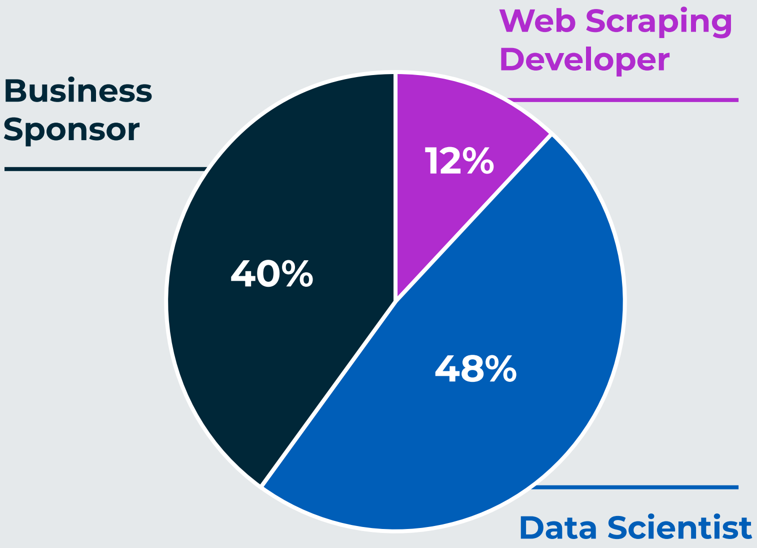
James Kehoe

Product Manager at Zyte

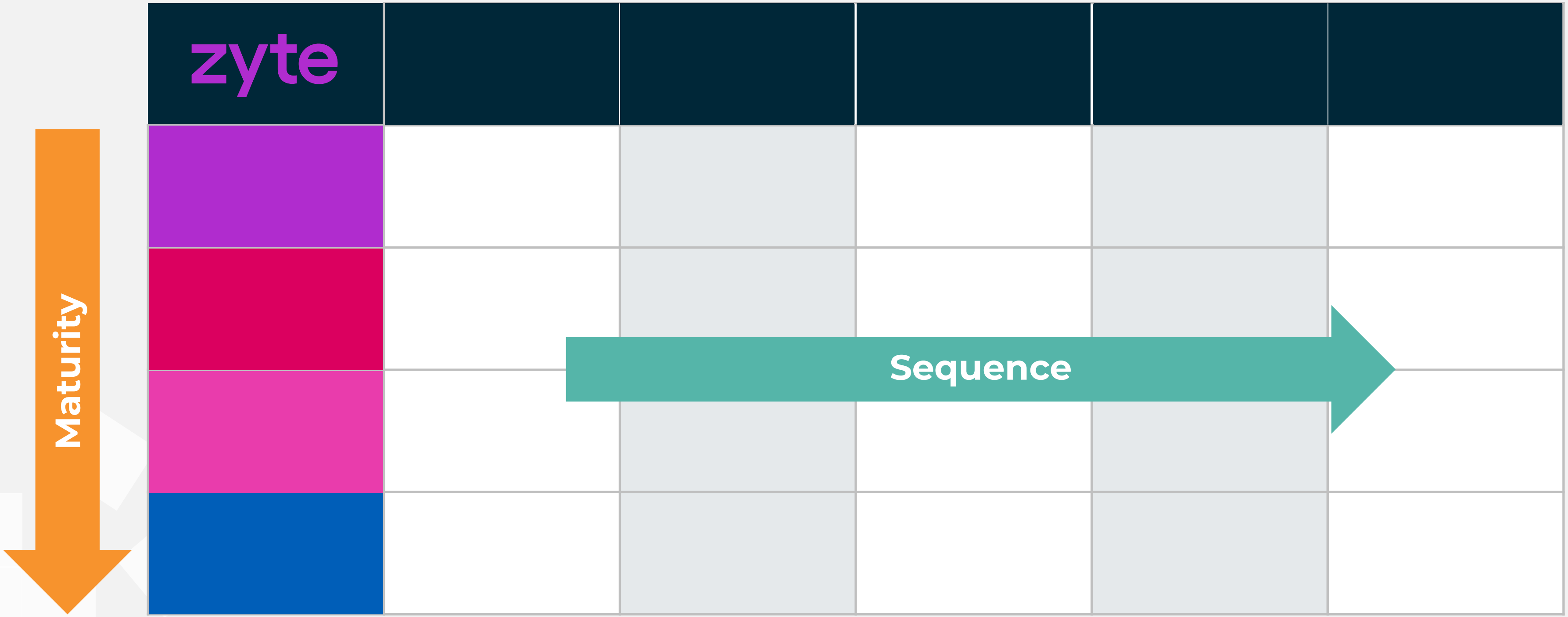
How we built the model

- **Interviewed internal teams**
  - 5,000 customers and 13B pages extracted a month

- **Interviewed 40+ industry representatives**
  - Business Sponsor - 40%
  - Data Scientist - 48%
  - Web Scraping Developer - 12%



Structure of the model



zyte	1. Creating the business case	2	3	4	5
Level-1: Ad-hoc	▪ **No documented business case** ▪ Single use case (e.g. pricing) ▪ Poor understanding of costs of web data ▪ Inappropriate web data success KPIs ▪ No commercial success KPIs				
Level-2: Opportunistic	▪ Very simple business case ▪ Just targets high profile sites ▪ **Scraped data not fully leveraged** ▪ Limited commercial success KPIs				
Level-3: Systematic	▪ Multiple data use cases ▪ Formal business cases ▪ **Data feed ROI considered** ▪ Comprehensive list of sites that complement each other. ▪ Extensive data schema				
Level-4: Proactive	▪ **Commercial governance for all data feeds** ▪ Prioritised backlog of data requests ▪ Regularly adding and removing data sources (i.e. sites, fields)				

Value of Business Cases

Prioritise investments

Investments with highest ROI get prioritised.

Increase probability of success

Written business cases force conversations that ensure those at the execution stage understand the objectives, saving time and increasing scope for innovate solutions.

Creates organisational memory

Good business cases become a reusable templates for future projects thereby reducing the effort of evaluating ideas.

zyte	1	2. Deploying resources	3	4	5
Level-1: Ad-hoc		▪ Partial resource allocation ▪ No external vendors			
Level-2: Opportunistic		▪ Temporary full time resources ▪ **Siloed resources** ▪ Leverages external vendors			
Level-3: Systematic		▪ Dedicated full time resources ▪ Multiple or Dominant vendor ▪ **Vendor as a business partner**			
Level-4: Proactive		▪ **Cross training & Career progression facilitated** ▪ Regular fallback testing ▪ QBR style engagements with vendors ▪ MSAs in place with multiple vendors			

Value of Deploying Resources

Ensures alignment with strategy

Organisations should prioritise investments in areas that give them a competitive advantage and are core to strategy. Resourcing not matching this strategy leads to uncertainty (e.g. hiring large team to do web scraping when this is not aligned with strategy)

Team sustainability over individual hires

“No one goes to college to maintain spiders!” To attract and retain talent you must have a structure, and challenges that support growth and development.

zyte	1	2	3. Ensuring compliance	4	5
Level-1: Ad-hoc			▪ Nothing considered ▪ **No legal review**		
Level-2: Opportunistic			▪ Considered but not acted on ▪ Legal review by generalist		
Level-3: Systematic			▪ Considered, documented and acted on ▪ **Dedicated legal specialist** ▪ Continuously monitor for new regulations		
Level-4: Proactive			▪ Documented, audited, and maintained ▪ Team of dedicated legal specialist working as partners ▪ **Continuously monitoring emerging case law**		

Value of Compliance

Sustainable business strategy

De-risk investments by ensuring data can be collected in a sustainable and compliant manner.

Protect your brand

Ensure your teams don’t inadvertently carry out activities that are illegal or do not comply with your corporate ethics.

zyte	1	2	3	4. Building feeds	5
Level-1: Ad-hoc				▪ **No antiban capabilities** ▪ Hacky spiders in different languages (python, JS, etc.) ▪ No shared utilities (e.g. text & html cleaners, formaters, etc.)	
Level-2: Opportunistic				▪ Consistent language ▪ Some tooling (e.g. SPM, SC, etc.) ▪ Some antiban capabilities	
Level-3: Systematic				▪ Consistent tooling ▪ **Sophisticated antiban approach (e.g. SMEs)** ▪ Single vendor/team ▪ Incremental crawling	
Level-4: Proactive				▪ **ML fall backs** ▪ Adaptive spiders ▪ Multiple vendors/solutions ▪ Dedicated antiban solutioning	

Value of Building Feeds

Build according to strategy

It’s easy to build a spider to get web data, but it’s much harder to build a monitorable, maintainable spider that functions in such as way as to ensure company strategy is realised (e.g. graceful failing, etc.).

Use constraints to your advantage

It takes time to get web data, use that to prioritise the data you need to scrape, and hence your crawling strategy (e.g. Product detail page weekly, price from Product list page daily, etc.).

	1	2	3	4	5. Maintaining feeds
Level-1: Ad-hoc					▪ Firefighting or disposable
Level-2: Opportunistic					▪ Builders = fixers (i.e. non-specialised) ▪ **No monitoring or major alerting**
Level-3: Systematic					▪ Manual QA ▪ Alerts & monitoring
Level-4: Proactive					▪ Manual & Automated QA ▪ Data feed health monitors ▪ **Monitoring for new or changed data fields**

Value of Maintaining Feeds

Use resource ratios

For example for every new data feed built there should be X FTEs available to maintain the feed (e.g. 0.02 FTEs).

Establish SLAs

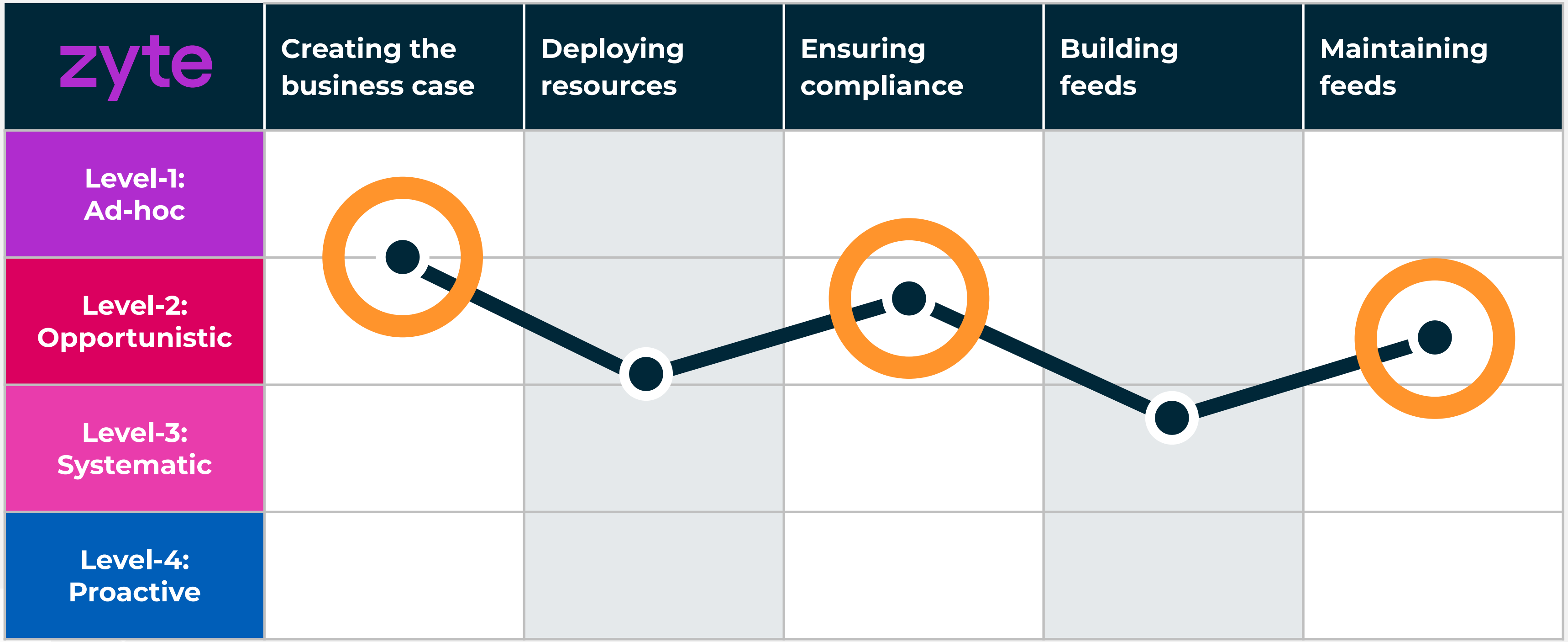
Response times, uptimes, and resolution times, all ensure appropriate monitoring and fixing resources are put in place.

Pro-active over reactive

Identify high risk time periods and ensure adequate resources are in place (e.g. around Black Friday).

zyte	1. Creating the business case	2. Deploying resources	3. Ensuring compliance	4. Building data feeds	5. Maintaining data feeds
Level-1: Ad-hoc	▪ No documented business case ▪ Single use case (e.g. pricing) ▪ Poor understanding of costs of obtaining web data ▪ Inappropriate web data success KPIs ▪ No commercial success KPIs	▪ Partial resource allocation ▪ No external vendors	▪ No legal review	▪ No antiban capabilities ▪ Hacky spiders in different languages (python, JS, etc.) ▪ No shared utilities (e.g. text & html cleaners, formaters, etc.)	▪ Firefighting or disposable
Level-2: Opportunistic	▪ Very simple business case ▪ Just targets high profile sites ▪ Scraped data not fully leveraged ▪ Limited commercial success KPIs	▪ Temporary full time resources ▪ Siloed resources ▪ Leverages external vendors	▪ Considered but not acted on ▪ Legal review by generalist	▪ Consistent language ▪ Some tooling (e.g. SPM, SC, etc.) ▪ Some antiban capabilities	▪ Builders = fixers (i.e. non-specialised) ▪ No monitoring or major alerting
Level-3: Systematic	▪ Multiple data use cases ▪ Formal business cases ▪ Data feed ROI considered ▪ Comprehensive list of sites that complement each other ▪ Extensive data scheme	▪ Dedicated full time resources ▪ Multiple or Dominant vendor ▪ Vendor as a business partner	▪ Considered, documented and acted on ▪ Dedicated legal specialist ▪ Continuously monitor for new regulations	▪ Consistent tooling ▪ Sophisticated antiban approach (e.g. SMEs) ▪ Single vendor/team ▪ Incremental crawling	▪ Manual QA ▪ Alerts & monitoring
Level-4: Proactive	▪ Commercial governance for all data feeds ▪ Prioritised backlog of data requests ▪ Regularly adding and removing data sources (i.e. sites, fields)	▪ Cross training & Career progression facilitated ▪ Regular fallback testing ▪ QBR style engagements with vendors ▪ MSAs in place with multiple vendors	▪ Documented, audited, and maintained ▪ Team of dedicated legal specialist working as partners ▪ Continuously monitoring emerging case law	▪ ML fall backs ▪ Adaptive spiders ▪ Multiple vendors/solutions ▪ Dedicated antiban solutioning	▪ Automated QA ▪ Data feed health monitors ▪ Monitoring for new or changed data fields

Average response of Interviewees



zyte	Creating the business case	Deploying resources	Ensuring compliance	Building feeds	Maintaining feeds
Level-1: Ad-hoc					
Level-2: Opportunistic					
Level-3: Systematic					
Level-4: Proactive					

Thank you