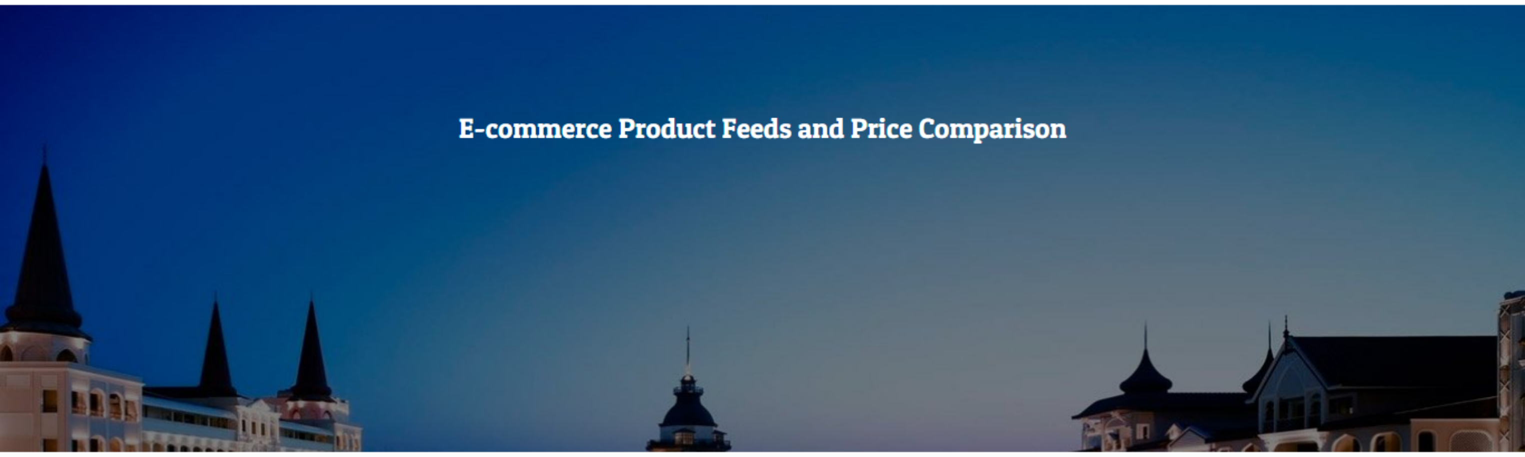
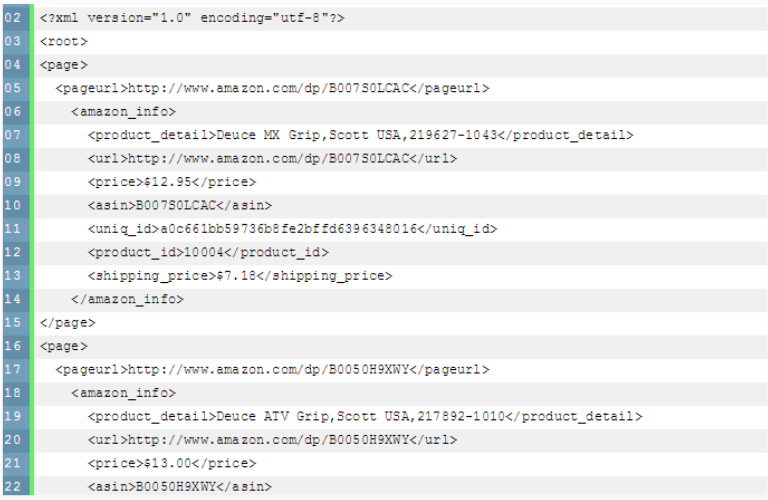


E-commerce Product Feeds and Price Comparison

Client: Data Analytics and Business research Shop for E-commerce and Retail

Challenge: With their expertise in analytics and research, the client was looking to include a data layer into its current set-up that would allow continuous free flowing feeds, free of "noise" so that the team could only focus on interesting approaches to analytics. They were interested in having easy access to a complete product listing from specific categories along with all the product specifications and prices listed together. The client previously had a data team that manually gathered data from various web sources but the results were limited and efforts were high. Even with the manual effort, structuring of the data in order to import it into their database was a challenge. Client was in need of clean data that could be uploaded into their DB in order to run the comparison engine and perform other monitoring activities.



```
<?xml version="1.0" encoding="utf-8"?>
<root>
<page>
  <pageurl>http://www.amazon.com/dp/B007S0LCAC</pageurl>
    <amazon_info>
      <product_detail>Deuce MX Grip,Scott USA,219627-1043</product_detail>
      <url>http://www.amazon.com/dp/B007S0LCAC</url>
      <price>$12.95</price>
      <asin>B007S0LCAC</asin>
      <uniq_id>a0c661bb59736b8fe2bffd6396348016</uniq_id>
      <product_id>10004</product_id>
      <shipping_price>$7.18</shipping_price>
    </amazon_info>
</page>
<page>
  <pageurl>http://www.amazon.com/dp/B0050H9XWY</pageurl>
    <amazon_info>
      <product_detail>Deuce ATV Grip,Scott USA,217892-1010</product_detail>
      <url>http://www.amazon.com/dp/B0050H9XWY</url>
      <price>$13.00</price>
      <asin>B0050H9XWY</asin>
```

Ecommerce sample data

The client provided us with the list of sources to be crawled and the data points required. The extraction was to be done on daily basis which meant fresh data sets have to be provided everyday. Our team set up crawlers to fetch the required data fields from the source sites provided by the client. This use case comes under our site specific crawl offering since the websites in the list had different structuring and design. The client needed the extracted data in CSV format and be uploaded to their S3 servers. The initial setup was complete in a few days and the crawlers started delivering data immediately. About 200 k records were delivered to the client during the first crawl.

The Solution: A crawler was set up that could extract product prices and specifications only for predefined categories in an automated manner on a daily basis. Based on the schema provided by the client, a template was created using which structuring of the data (extraction) would occur. The final data was delivered in an XML format via the Data API on a daily basis without any manual intervention from either side. Each record within a dataset had all details i.e. product name, product price, availability status, short and long descriptions, all image URL's, SKU, dimensions, category, brand, source and the source URL from where it was fetched.

Benefits:

- Any changes within the source sites were taken care of and clients were abstracted from such issues
- Any changes with respect to schema was done as requested
- Other categories could be added as per changing requirements
- Productivity increased since the data team could work on other projects. Client expanded into other verticals
- Low turnaround time of data improved the ability to market client's services and capabilities
- Value addition from the project was 50 times the spend
- Data quality levels had increased alarmingly without any time investment from the team