

no I will not talk about

Google™

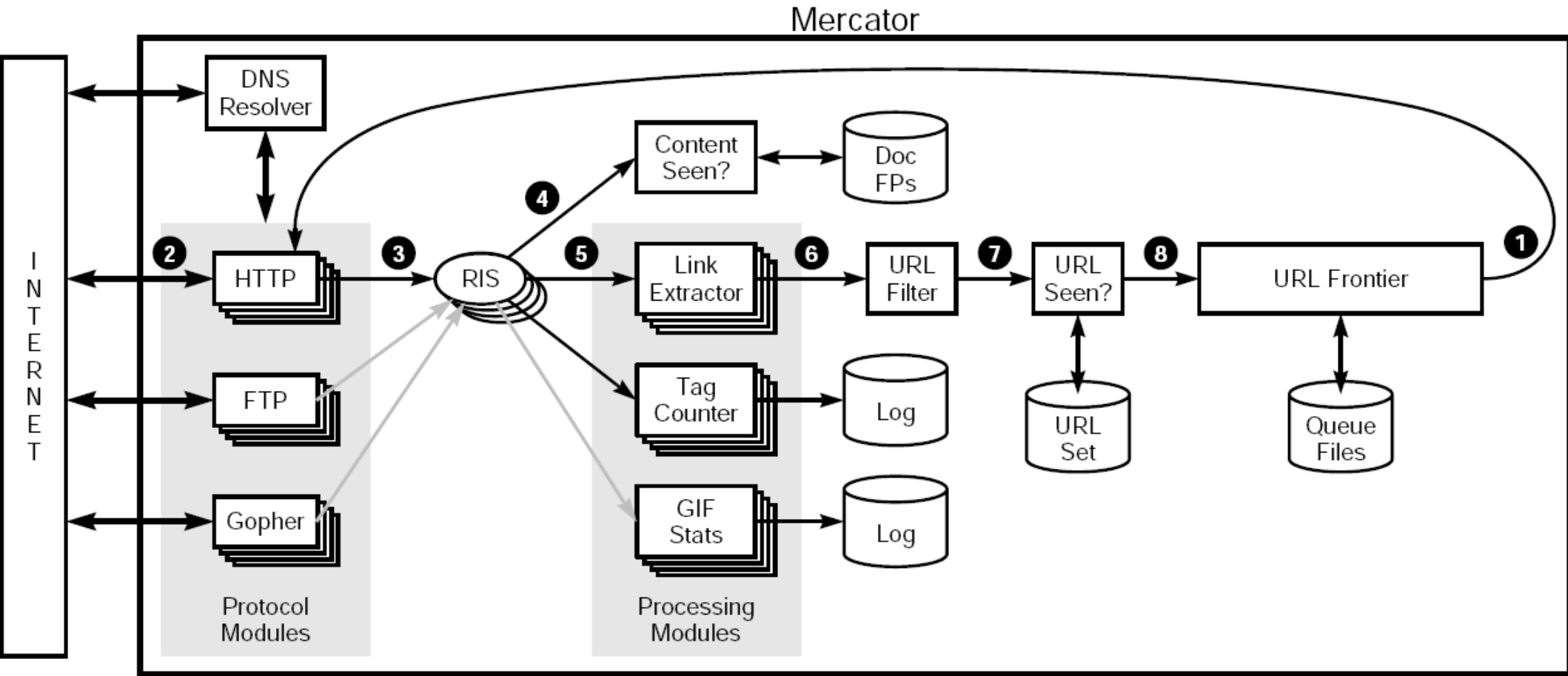
# The Mercator Web Crawler and a peek at OpenSearch

[Nirmal Thacker]

# Introduction

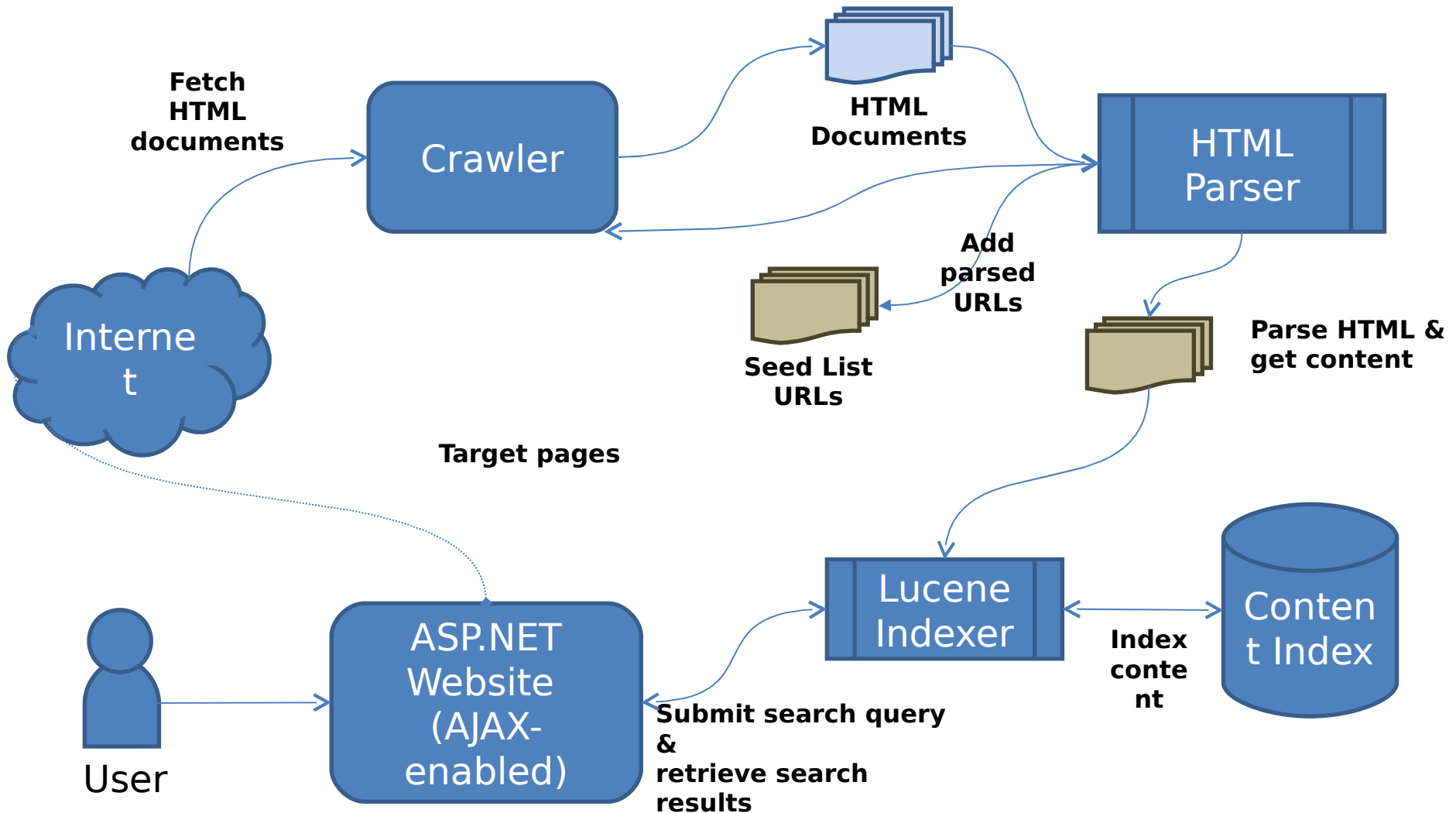
- Crawlers aka web-spider aka web-robots.
- The process – web crawling or spidering.
- Crawling is the first step to an effective search [google story]
- With the current size of the web the issues important to a crawler are:
  - page freshness
  - cache what on the basis of what
  - what kind of i/o would you need
  - how many machines- work in parallel? if so- what about duplication- a crawler seems to be a challenge in distributed systems!
- Mercator is a scalable, extensible web crawler written entirely in Java.

# Mercator Architecture

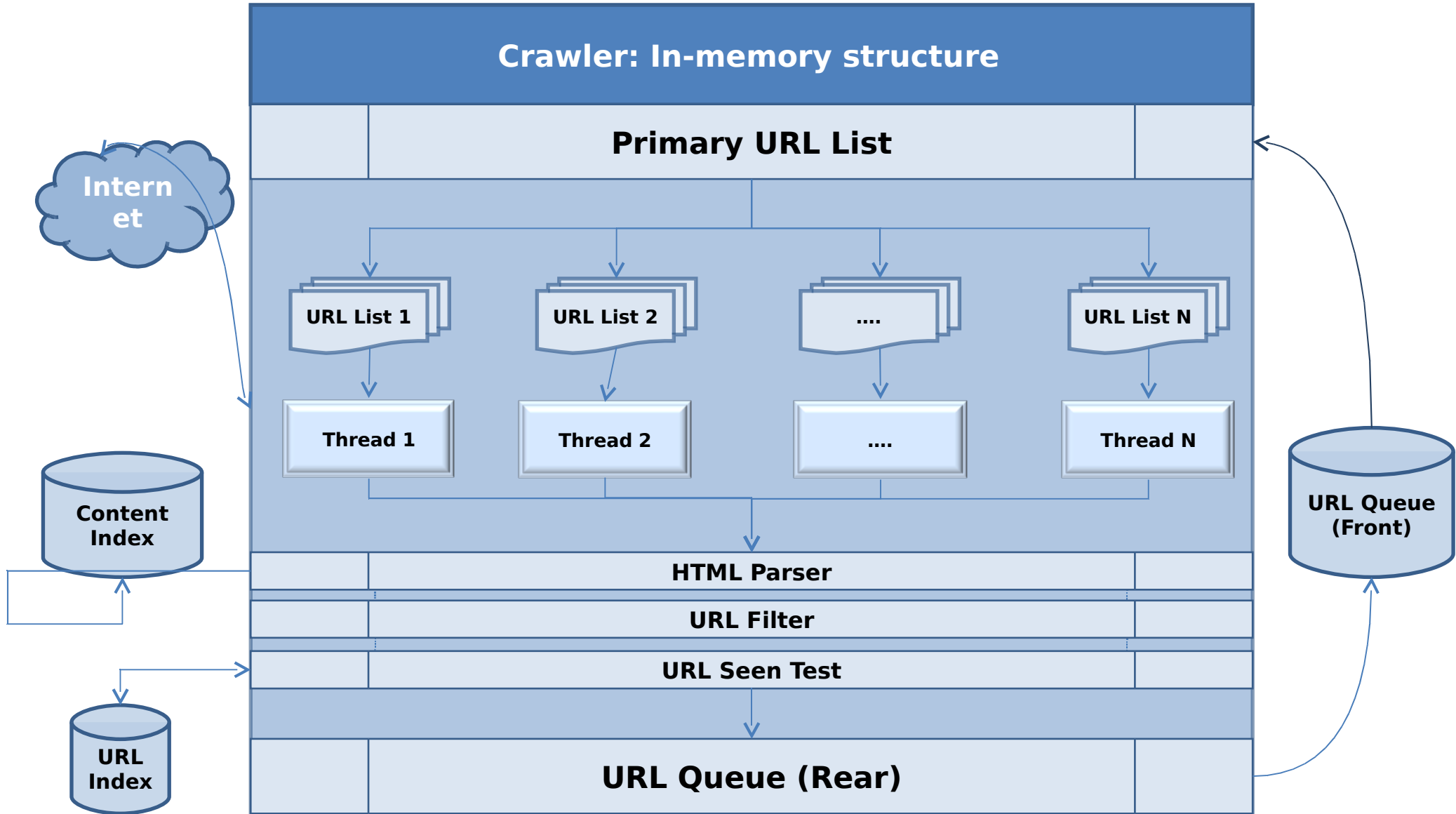


# Mercator in .Net [with permissions from Sunil

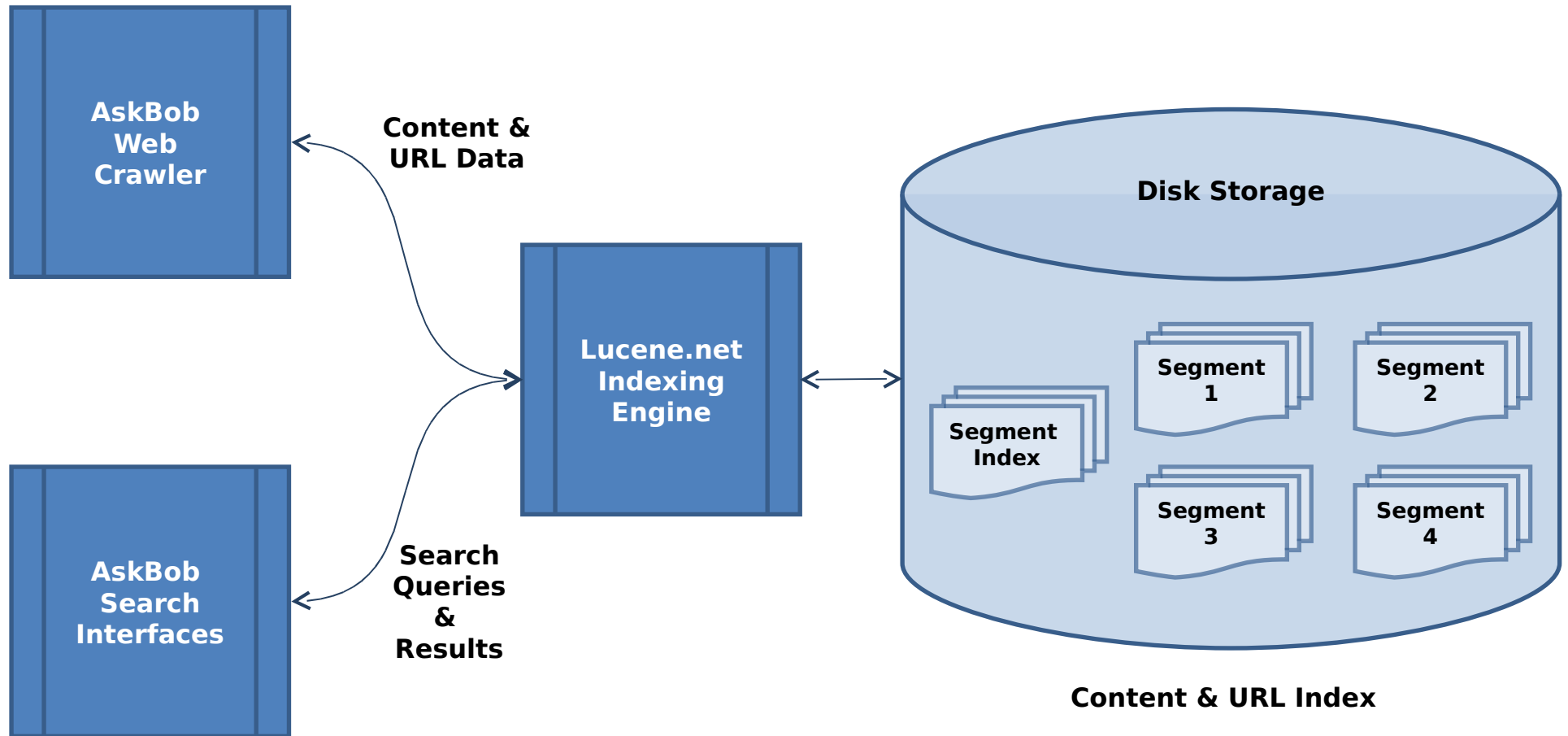
Jagadish <http://suniljagadish.blogspot.com/> & Sharath MS-  
<http://www.sharathms.com/>]



# Another look at the architecture



# An indexer in .Net- Lucene



# AskBob Demo

# OpenSearch [<http://www.opensearch.org/>]

- Created by A9 – an Amazon company but its open and anyone can use it.
- One algorithm is not enough for all the information available.
- You may need specific domain information/collaborated information/design an info-spec for your database itself
- A description document can describe your search interface
- Response elements can allow 3<sup>rd</sup> party apps to use your search engine the way you want it to be uniquely used

# Questions?

# More at:

- The Mercator Website:<http://mercator.comm.nsdlib.org/>
- Paper: Mercator: A Scalable, Extensible Web Crawler (1999)- Allan Heydon, Marc Najork [on the class website]
- <http://www.opensearch.org/>
- Some OpenSearch clients you can try out:
  - <http://a9.com/>
  - <http://osfeed.com/>
  - <http://tagjag.com/>