Please find my comments to the requirements remarked in blue!!

First of all I want to say congratulation to your description of the requirements . The requirements are very clear for me, lot of other posted jobs are very unstructured.

Requirement:

1. We have a about million+ URL seeds to crawl from. We would like to build a heavy-duty crawler like the search engine use. (Though we would be limiting the crawling to specific domains for the first release.)
How should work the crawler: Should the crawler handle pages recursive base on the root domain without links depths or should the crawler handling extract links with depth config parameter? Should we design the database schema with intergrated page rankings and plugin interface for data analyses ? or you have already a database and crawler schema ?

2. The crawler must look for specific meta-tags and some outgoing URL formats, before extracting the page content and store it in a DB for post-analysis.

3. The crawler must run as a service repeatedly as configured.
Do your mean as a service with configuration parameter or as a cron job with arguments on the command line?

4. Require simple dashboard to monitor the crawler performance.
The crawler will delivered with integrated reports interface

Please bid if you had done similar work and if you can leverage a stable framework or reuse some code base. Also share some performance numbers in terms of concurrency, how many pages you can crawl in an hour or day etc.

The average is depending of the E2C server capacity and if crawler need to interpret javascript.. I you use m3.2xlarge Amazon servers and you need to interpret javascript you can reach capacity of : 2-3 million pages handling / day in one E2C Instance. Please note that this result could be improve if max parallel jobs setting change from 1000 to 10000.

Please take a look on my example : https://github.com/ovntatar/Elance-Bala-V-Airomo

1 E2C Avarage =

5 min / 10000 pages
10 min 20 000 pages
60 min(1 hour) 120 000 pages
120 000 x 24 hours = 2 880 000 pages / day in one E2C instance

The efficiently is dependent of the HTTP server handling. If serve use encryption, forwarding etc.. .

The serviceice will be hosted from Amazon EC2.

You want to respect client bandwidth and robots.txt ? My advantage to my competetation is (I hope that this addition service from 3zele.de would help to make your desicion fo us) - I can generate the crawler to run anonym to avoid any judge topics and respect server privacy ;-)